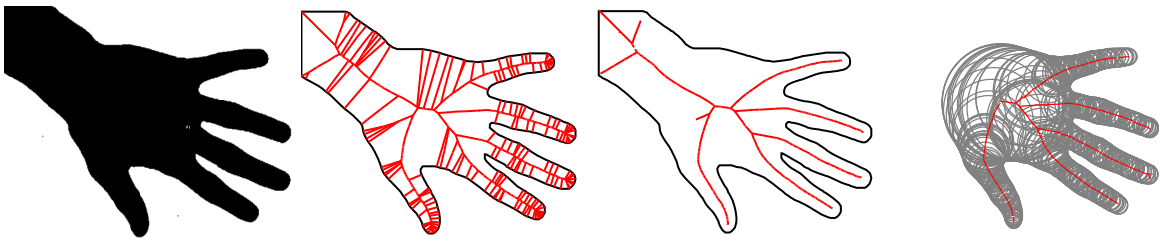
ISSN 2223-3792

Машинное обучение и анализ данных

2014 год **Том 1, номер 9**



(а) (б) (в) (г)

Машинное обучение и анализ данных

Journal of Machine Learning and Data Analysis

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области теоретических основ информатики и её приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на английском и русском языках.

Журнал включен в российский индекс научного цитирования РИНЦ. Информация о цитировании статей находится на сайте Российского индекса научного цитирования. ISSN 2223-3792, номер свидетельства о регистрации ЭЛ № ФС 77-55486.

• Архив журнала http://www.ccas.ru/jmlda/
• Новостной сайт http://jmlda.org/
• Электронная система подачи статей http://jmlda.org/papers/

Тематика журнала:
— классификация, кластеризация, регрессионный анализ,
— алгебраический подход к проблеме синтеза корректных алгоритмов,
— многомерный статистический анализ,
— выбор моделей и сложность,
— предсказательное моделирование,
— статистическая теория обучения,
— методы прогнозирования временных рядов,
— методы обработки и распознавания сигналов,
— методы оптимизации в задачах машинного обучения и анализа данных,
— методы визуализации данных,
— обработка и распознавание речи и изображений,
— анализ и понимание текста,
— информационный поиск,
— прикладные задачи анализа данных.

Вычислительный центр Российской академии наук
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2014

Содержание

Двухкомпонентная функция качества кластеризации множества элементов, представленных парными сравнениями*

С. Д. Двоенко
dsd@tsu.tula.ru
Тульский государственный университет, Россия, Тула, пр. Ленина, 92

Рассмотрены варианты известного алгоритма k-средних, в которых не требуется вычислять собственно средние по кластерам. В новых версиях алгоритма k-средних выполняются перестановки на матрице парных сравнений так, что в случае помещения анализируемого множества объектов в признаковое пространство достигается тот же самый результат кластеризации. Рассмотрена новая двухкомпонентная целевая функция качества кластеризации как минимизируемая комбинация внутрикластерных дисперсий (квадратов расстояний) с близостью кластеров между собой или, в двойственной формулировке, как максимизируемая комбинация внутрикластерных близостей с дисперсией (квадратами расстояний) между кластерами. Показано, что качество кластеризации удается улучшить по сравнению с обычным критерием качества кластеризации.

Ключевые слова: кластер; k-средних; расстояние; близость; беспризнаковый

Bi-partial objective function for clustering a set of elements in terms of pairwise comparisons*

S. D. Dvoenko
Tula State University, Russia, Tula, Lenin Ave., 92

Background: In a featureless case, a set of objects is represented only by results of pairwise mutual comparisons in the form of a distance, similarity, or kernel-based matrix. Nevertheless, the cluster centers can be implicitly represented by its distances to other objects without the feature space itself.

Methods: The present author proposes k-means clustering without computations of cluster centers at all. This novel procedure, referred to as the k-meanless clustering, makes permutations on the similarity or distance square matrix resulting in the same clustering for both featureless and feature-based cases. In addition, new bi-partial objective function combines intracluster distances with intercluster similarities and needs to be minimized or in the dual form combines intracluster similarities with intercluster distances and needs to be maximized.

Results: Based on bi-partial approach, the clustering quality can be improved relative to the usual objective function.

Concluding Remarks: The k-means idea is very popular in the form of many heuristic aggregating procedures where cluster centers cannot be explicitly presented. Therefore, they are only suboptimal versions of the k-means. The proposed k-meanless clustering is the correct version of them.

Keywords: cluster; k-means; distance; similarity; featureless

*Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00010.

Погружение элементов множества в метрическое пространство

В задаче кластер-анализа объекты $\omega_i \in \Omega$, $i = 1, \ldots, N$ обычно представлены как векторы $\mathbf{x}_i = (x_{i1}, \ldots x_{in})^{\mathrm{T}}$ в n-мерном пространстве признаков и образуют матрицу данных $X(N, n)$. В соответствии с гипотезой компактности объекты образуют локальные сгущения в виде K кластеров (классов, таксонов).

Хорошо известные алгоритмы типа k-средних [1] основаны на идее несмещенного разбиения [2]. В соответствии с ней каждый кластер Ω_k, $k = 1, \ldots, K$, представлен своим «представителем» $\tilde{\mathbf{x}}_k$, а центр кластера представлен средним $\bar{\mathbf{x}}_k$.

Если окажется, что для всех кластеров представители и центры совпадают $\tilde{\mathbf{x}}_k = \bar{\mathbf{x}}_k$, то получена несмещенная кластеризация, а противном случае – смещенная. Тогда необходимо назначить центры (средние объекты) в качестве новых представителей, заново расклассифицировать объекты по минимуму расстояния до представителей и вычислить новые центры кластеров.

В случае, когда признаковое пространство нам недоступно, средний объект $\omega(\bar{\mathbf{x}}_k)$ не представлен в матрице расстояний $D(N, N)$ как центр соответствующего кластера. Поэтому обычно в качестве эвристических агрегирующих процедур применяют некорректные версии алгоритма k-средних, где вместо центра кластера $\bar{\omega}_k$ в таком качестве используют объект, ближайший ко всем остальным в кластере. Тогда в общем случае при выполнении всех условий $\tilde{\omega}_k = \bar{\omega}_k$ может быть получена смещенная кластеризация, так как при погружении данного множества в соответствующее признаковое пространство окажется, что центр кластера $\mathbf{x}(\bar{\omega}_k)$ может не совпадать со средним объектом $\bar{\mathbf{x}}_k$.

Кластеризация по расстояниям до центров кластеров

Как известно, среднее арифметическое, используемое в качестве центра кластера, минимизирует его дисперсию и как результат дисперсию всей кластеризации [1].

Дисперсия кластера представлена квадратами отклонений объектов от центра кластера, т. е. квадратами соответствующих расстояний:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\mathbf{x}_i,\ \bar{\mathbf{x}}_k)\,.$$

Очевидно, что данный критерий минимизирует среднее квадратов расстояний до центра кластера и средневзвешенную дисперсию кластеризации в целом:

$$J(K) = \frac{1}{N} \sum_{k=1}^{K} N_k \sigma_k^2 = \sum_{k=1}^{K} \frac{N_k}{N} \sigma_k^2\,.$$

В отсутствие признаков средние объекты $\bar{\omega}_k$ обеспечивают несмещенную кластеризацию, также минимизируя дисперсии кластеров:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\omega_i,\ \bar{\omega}_k)$$

и значение критерия $J(K)$ в целом.

Если множество Ω будет помещено в соответствующее пространство признаков, где объекты $\mathbf{x}(\bar{\omega}_k)$ и $\bar{\mathbf{x}}_k$ совпадут, то два критерия:

$$J^X(K) = \min_{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_K} J(K);\quad J^D(K) = \min_{\bar{\omega}_1, \ldots, \bar{\omega}_K} J(K)$$

окажутся одинаковыми $J^X(K) = J^D(K)$. Очевидно, что $J^D(K) \geqslant J^X(K)$ в общем случае. Построим алгоритм для получения несмещенной кластеризации.

Для некоторого элемента $\omega_l \in \Omega$, взятого как начало координат, и пары $\omega_i,\ \omega_j$ их скалярное произведение $c_{ij} = (d_{li}^2 + d_{lj}^2 - d_{ij}^2)/2$ вычисляется на основе расстояний $d_{pq} = d(\omega_p,\ \omega_q)$, где $c_{ii} = d_{li}^2$ при $i = j$.

Следовательно, элементы главной диагонали матрицы $C_l(N, N)$ представляют собой квадраты расстояний от начала координат $\omega_l \in \Omega$ до остальных объектов. Удобно [3] поместить начало координат в центр тяжести множества $\omega_i \in \Omega,\ \ i = 1, \ldots, N$.

Как показано в [4, 5, 6, 7], можно немедленно доказать, что центр кластера $\bar{\omega}_k$ будет представлен своими расстояниями до остальных объектов $\omega_i \in \Omega,\ \ i = 1, \ldots, N$ без необходимости восстановления неизвестного нам признакового пространства, где N_k — число объектов в кластере Ω_k:

$$d^2(\omega_i,\ \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2;\ \omega_p,\ \omega_q \in \Omega_k\ ,$$

где дисперсия кластера вычисляется как:

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) = \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\ . \quad (1)$$

Известно, что алгоритм k-средних может быть представлен в различных вариантах в соответствии со способами пересчета средних в признаковом пространстве. Представим данный алгоритм для расстояний в нужном нам виде, где пересчет центров выполняется сразу после очередного переноса.

Алгоритм 1:

Шаг 0. Взять в качестве центров $\bar{\omega}_k^0,\ \ k = 1, \ldots, K$, например, K наиболее удаленных друг от друга объектов и назначить их представителями $\tilde{\omega}_k^0,\ \ k = 1, \ldots, K$.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $d(\omega_i,\ \bar{\omega}_k^s) \leqslant d(\omega_i,\ \bar{\omega}_j^s)$, где $j = 1, \ldots, K,\ j \neq k$.

2. Пересчитать, если требуется, центры $\bar{\omega}_k^s,\ \ k = 1, \ldots, K$ и представить их своими расстояниями до всех объектов $d(\omega_i,\ \bar{\omega}_k^s),\ \ i = 1, \ldots, N$.

3. Переместить следующий $i = i + 1$ объект ω_i.

4. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация, где $\tilde{\omega}_k^s = \bar{\omega}_k^s$, $k = 1, \ldots, K$, иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$ и перейти к следующему шагу $s = s + 1$.

Кластеризация перестановками без центров кластеров

Заметим, что можно также вычислить среднее квадратов расстояний между объектами в кластере. С учетом расстояний до себя получим выражение:

$$\eta_k' = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(\mathbf{x}_i - \mathbf{x}_j)^2 = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\mathbf{x}_i,\ \mathbf{x}_j)\ .$$

Из (1) для σ_k^2 немедленно следует, что $\eta'_k = 2\sigma_k^2$. Введем обозначение $\eta_k = \eta'_k/2 = \sigma_k^2$, где

$$\eta_k = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\omega_i,\ \omega_j)\ .$$

Следовательно, для всех кластеров минимизация взвешенных квадратов расстояний между объектами в кластерах приводит к минимизации средневзвешенной дисперсии кластеризации в целом:

$$\tilde{J}(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\eta_k = \sum_{k=1}^{K}\frac{N_k}{N}\eta_k\ .$$

Следовательно, критерии $\tilde{J}(K)$ и $J(K)$ совпадают $\tilde{J}(K) = J(K)$. Если множество Ω будет помещено в соответствующее пространство признаков, то объекты $\mathbf{x}(\bar{\omega}_k)$ и $\bar{\mathbf{x}}_k$ совпадут, где два критерия:

$$\tilde{J}^X(K) = \min_{\Omega_1,\dots,\Omega_K\in X}\tilde{J}(K) \text{ и } \tilde{J}^D(K) = \min_{\Omega_1,\dots,\Omega_K\in D}\tilde{J}(K)$$

также совпадут $\tilde{J}^X(K) = \tilde{J}^D(K)$. Очевидно, что в общем случае $\tilde{J}^D(K) \geqslant \tilde{J}^X(K)$.

Построим кластеризацию без центров. Очевидно, что такая кластеризация должна быть несмещенной, если для нее вычислить центры кластеров. Эквивалентная модификация алгоритма k-средних, рассмотренного выше, имеет следующий вид.

Алгоритм 2:

Шаг 0. Взять в качестве подмножеств $\Omega_k^0,\quad k = 1,\dots,K$, например, K наиболее компактных в некотором смысле подмножеств.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$ и принять $\tilde{J}^s(K) = \tilde{J}_k^s(K)$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$, выполнено условие $\tilde{J}_k^s(K) < \tilde{J}_j^s(K),\ j = 1,\dots K,\ j \neq k$.

2. Переместить следующий $i = i + 1$ объект ω_i.

3. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация. Иначе перейти к следующему шагу $s = s + 1$.

Кластеризация по близостям

Положительно полуопределенная матрица близостей $S(N,N)$ с элементами $s_{ij} = s(\omega_i,\ \omega_j) \geqslant 0$ может рассматриваться как матрица скалярных произведений в метрическом пространстве размерности не выше N. Относительно некоторой точки $\omega_k \in \Omega$, взятой как начало координат, где $s_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2$, $s_{ii} = d_{ki}^2$, расстояния определяются как $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$.

В данном случае центр кластера $\bar{\omega}_k$ может быть представлен своими близостями к остальным объектам $\omega_i \in \Omega,\ \ i = 1,\dots N$, где N_k – число объектов в кластере Ω_k:

$$s(\omega_i,\bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} s_{ip};\ \omega_p \in \Omega_k,\ \omega_i \in \Omega,\ \ i = 1,\dots,N\ .$$

Компактность кластера может быть представлена как средняя близость центра к остальным объектам в кластере:

$$\delta_k = \frac{1}{N_k}\sum_{i=1}^{N_k} s(\omega_i,\bar{\omega}_k) = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{p=1}^{N_k} s_{ip};\ \omega_i,\omega_p \in \Omega_k\ .$$

Несмещенная кластеризация минимизирует дисперсии кластеров σ_k^2 и максимизирует их компактности δ_k, где с учетом $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$ получим:

$$\sigma_k^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(s_{ii} + s_{jj} - 2s_{ij}) = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} s_{ij} = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k \ .$$

Тогда для всех кластеров получим:

$$J(K) = \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k\right) = \frac{1}{N}\sum_{i=1}^{N} s_{ii} - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k = c - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k \ .$$

Обозначим средневзвешенную компактность кластеризации, которую следует максимизировать, в виде нового функционала:

$$I(K) = \sum_{k=1}^{K}\frac{N_k}{N}\delta_k, \text{ где } \quad I(K) = c - J(K). \tag{2}$$

Немедленно получим две модификации алгоритма k-средних для близостей: с вычислением центров кластеров и без них. Построим алгоритм кластеризации с вычислением центров.

Алгоритм 3:

Шаг 0. Взять в качестве центров $\bar{\omega}_k^0, \quad k = 1, \ldots K$, например, K наименее близких друг к другу объектов и назначить их представителями $\tilde{\omega}_k^0, \ \ k = 1, \ldots, K$.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $s(\omega_i, \ \bar{\omega}_k^s) \geqslant s(\omega_i, \ \bar{\omega}_j^s)$, где $j = 1, \ldots, K, \ j \neq k$.

2. Пересчитать, если требуется, центры $\bar{\omega}_k^s, \ k = 1, \ldots K$ и представить их своими близостями ко всем объектам $s(\omega_i, \ \bar{\omega}_k^s), \ \ i = 1, \ldots, N$.

3. Переместить следующий $i = i + 1$ объект ω_i.

4. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация, где $\tilde{\omega}_k^s = \bar{\omega}_k^s, \ k = 1, \ldots, K$, иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$ и перейти к следующему шагу $s = s + 1$.

Построим алгоритм кластеризации по близостям без центров.

Алгоритм 4:

Шаг 0. Взять в качестве подмножеств $\Omega_k^0, \quad k = 1, \ldots, K$, например, K наиболее компактных в некотором смысле подмножеств.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$ и принять $I^s(K) = I_k^s(K)$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$, выполнено условие $I_k^s(K) > I_j^s(K), \ j = 1, \ldots, K, \ j \neq k$.

2. Переместить следующий $i = i + 1$ объект ω_i.

3. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация. Иначе перейти к следующему шагу $s = s + 1$.

Двухкомпонентная целевая функция качества кластеризации

Легко увидеть, что для всех вариантов классического критерия качества кластеризации как для признакового пространства, так и в случае только парных сравнений, справедливо общее свойство: минимизируя разброс объектов в кластерах (или максимизируя

«плотность» кластеров в двойственной формулировке) мы никак не управляем разбросом центров кластеров. Очевидно, что в общем случае, делая кластеры более плотными, желательно еще попробовать и отдалить их друг от друга, насколько это возможно.

Реализация такого подхода [8, 9] приводит к построению двухкомпонентной целевой функции качества кластеризации. При построении такой функции возникает проблема масштабирования двух ее частей: одна из них отвечает за внутриклассовые характеристики, а другая – за межклассовые. Это приводит в общем случае к необходимости выбора соответствующих шкал измерений, зависящих от интерпретации понятия «кластер» и к согласованию их путем поиска оптимальной линейной комбинации.

В нашем случае, в отличие от рассмотренного подхода, задача оказывается проще. А именно: нам не требуется поиск согласованных шкал измерений для внутри- и межклассовых характеристик качества кластеризации, так как в беспризнаковом подходе расстояния и близости между элементами множества представлены в одном и том же метрическом пространстве, пусть даже и неизвестном. В этом случае потребуется просто определить масштаб влияния дополнительной части на общее значение критерия качества путем подбора соответствующего коэффициента.

Рассмотрим снова критерий $J(K)$ и его вариант $\tilde{J}(K)$ при отсутствии явно вычисленных центров кластеров. С учетом (1) для дисперсии η_k кластера Ω_k получим:

$$\tilde{J}(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\eta_k = \frac{1}{N}\sum_{k=1}^{K}\frac{N_k}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 = \frac{1}{2N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 \ .$$

Согласно (2), в двойственной формулировке для критерия $I(K)$ также получим:

$$I(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\delta_k = \frac{1}{N}\sum_{k=1}^{K}\frac{N_k}{N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} s_{pq} = \frac{1}{N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} s_{pq} \ .$$

Сначала рассмотрим двухкомпонентную целевую функцию качества кластеризации $\tilde{J}_\delta(K) = \tilde{J}(K) + \delta(K)$, которую нужно минимизировать, при комбинировании внутрикластерных дисперсий $\tilde{J}(K)$ с близостью между кластерами $\delta(K)$.

Как и ранее для кластеров, рассмотрим центр всего множества и обозначим его как новый элемент $\bar{\omega}_0$, который представим своими близостями в данном случае не ко всем элементам множества, а только к центрам других кластеров:

$$s(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K}\sum_{p=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_p),\ k = 1, \ldots, K \ .$$

Компактность множества центров кластеров, которую будем рассматривать как близость между кластерами $\delta(K)$, может быть представлена как средняя близость центра всего множества $\bar{\omega}_0$ к центрам кластеров $\bar{\omega}_k,\ k = 1, \ldots, K$:

$$\delta(K) = \frac{1}{K}\sum_{k=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_0) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{K}\sum_{l=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_l) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_l) \ .$$

Определим близости $s(\bar{\omega}_k,\ \bar{\omega}_l)$ центров кластеров друг к другу. Центр кластера $\bar{\omega}_k$ представлен своими близостями ко всем остальным объектам $\omega_i \in \Omega$ и, в частности, к объектам из другого кластера $\omega_i \in \Omega_l$. Рассмотрим среднюю близость объектов из другого

кластера $\omega_i \in \Omega_l$ к центру данного кластера $\bar{\omega}_k$:

$$s(\Omega_l,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l} s(\omega_i,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l}\frac{1}{N_k}\sum_{p=1}^{N_k} s_{ip} = \frac{1}{N_l N_k}\sum_{i=1}^{N_l}\sum_{p=1}^{N_k} s_{ip},\ \omega_p \in \Omega_k\ .$$

Очевидно, что $s(\Omega_l,\ \bar{\omega}_k) = s(\Omega_k,\ \bar{\omega}_l)$, так как $s_{ij} = s_{ji}$. Следовательно, можно использовать обозначения $s(\Omega_l,\ \bar{\omega}_k) = s(\Omega_k,\ \bar{\omega}_l) = s(\Omega_l,\ \Omega_k) = s(\bar{\omega}_l,\ \bar{\omega}_k)$ для парной близости между кластерами. В итоге близость между всеми кластерами выражается следующим образом:

$$\delta(K) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} s_{pq},\quad \text{где}\quad \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ .$$

Теперь рассмотрим двухкомпонентную целевую функцию качества кластеризации $I_\sigma(K) = I(K) + \sigma^2(K)$, которую нужно максимизировать, при комбинировании внутрикластерных близостей $I(K)$ с межкластерной дисперсией $\sigma^2(K)$.

Рассмотрим центр всего множества как объект $\bar{\omega}_0$, который представим своими расстояниями до центров других кластеров. Как показано в [4, 5, 6] и согласно (1) получим:

$$d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q),\ k = 1, \ldots, K\ .$$

Дисперсия множества центров кластеров $\sigma^2(K)$ может быть представлена как среднее квадратов расстояний от центра всего множества $\bar{\omega}_0$ до центров кластеров $\bar{\omega}_k,\ k = 1, \ldots, K$:

$$\sigma^2(K) = \frac{1}{K}\sum_{k=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_0) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q)\right) =$$

$$= \frac{1}{K^2}\sum_{k=1}^{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q) = \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q)\ .$$

Определим расстояния $d^2(\bar{\omega}_k,\ \bar{\omega}_l)$ между центрами кластеров. Центр кластера $\bar{\omega}_k$ представлен своими расстояниями до остальных объектов $\omega_i \in \Omega$ и, в частности, до объектов из другого кластера $\omega_i \in \Omega_l$. Рассмотрим среднее квадратов расстояний объектов из другого кластера $\omega_i \in \Omega_l$ до центра данного кластера $\bar{\omega}_k$:

$$d^2(\Omega_l,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l} d^2(\omega_i,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) =$$

$$= \frac{1}{N_l N_k}\sum_{i=1}^{N_l}\sum_{p=1}^{N_k} d_{ip}^2\ - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2,\ \ \omega_p,\ \omega_q \in \Omega_k\ .$$

Аналогично получим:

$$d^2(\Omega_k,\ \bar{\omega}_l) = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\omega_i,\ \bar{\omega}_l) = \frac{1}{N_k N_l}\sum_{i=1}^{N_k}\sum_{p=1}^{N_l} d_{ip}^2\ - \frac{1}{2N_l^2}\sum_{p=1}^{N_l}\sum_{q=1}^{N_l} d_{pq}^2,\ \omega_p,\ \ \omega_q \in \Omega_l\ .$$

Легко увидеть, что в общем случае $d^2(\Omega_l,\ \bar{\omega}_k) \neq d^2(\Omega_k,\ \bar{\omega}_l)$ из-за различных внутрикластерных дисперсий кластеров Ω_l и Ω_k.

Рассмотрим величину $d^2(\Omega_l,\ \Omega_k) = (1/2)\,(d^2(\Omega_l,\ \bar{\omega}_k) + d^2(\Omega_k,\ \bar{\omega}_l))$ как расстояние между двумя множествами. Очевидно, что при $d_{ij} = d_{ji}$ получим:

$$d^2(\Omega_l,\ \Omega_k) = \frac{1}{N_k N_l}\sum_{i=1}^{N_k}\sum_{p=1}^{N_l} d_{ip}^2 - \frac{1}{2}\left(\frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 + \frac{1}{2N_l^2}\sum_{s=1}^{N_l}\sum_{t=1}^{N_l} d_{st}^2\right),$$

где $\omega_p,\ \omega_q \in \Omega_k$ и $\omega_s,\ \omega_t \in \Omega_l$. В этом случае можно также ввести обозначение:

$$d^2(\bar{\omega}_l,\ \bar{\omega}_k) = d^2(\Omega_l,\ \Omega_k) = \frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2 - \frac{1}{2}(\sigma_k^2 + \sigma_l^2)\ .$$

Тогда межкластерная дисперсия выражается следующим образом:

$$\sigma^2(K) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2 - \frac{1}{2}(\sigma_k^2 + \sigma_l^2)\right),\ \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ .$$

Если внутрикластерные дисперсии не учитывать, то окажется, что $d^2(\Omega_l,\ \bar{\omega}_k) = d^2(\Omega_k,\ \bar{\omega}_l)$, так как $d_{ij} = d_{ji}$. Тогда для парных расстояний между центрами кластеров также можно использовать обозначения $d^2(\Omega_l,\ \bar{\omega}_k) = d^2(\Omega_k,\ \bar{\omega}_l) = d^2(\Omega_l,\ \Omega_k) = d^2(\bar{\omega}_l,\ \bar{\omega}_k)$. В этом случае межкластерная дисперсия выражается следующим образом:

$$\sigma^2(K) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2\ ,\ \text{где}\quad \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ .$$

Очевидно, что при $K = N$ в обоих случаях мы получим дисперсию всего множества, так как дисперсии одноэлементных кластеров являются нулевыми.

Таким образом, в случае двухкомпонентной функции качества кластеризации применяются те же алгоритмы k-средних (варианты 2 и 4), но только для критериев $\tilde{J}_\delta(K)$ и $I_\sigma(K)$.

Вычисление компонент целевой функции

Очевидно, что в двухкомпонентных целевых функциях $\tilde{J}_\delta(K)$ и $I_\sigma(K)$ необходимо одновременно использовать представление элементов множества как расстояниями, так и близостями между ними. Пусть задана матрица близостей $S(N,N)$ с элементами $s_{ij} = s(\omega_i,\ \omega_j) \geqslant 0$. Тогда элементы матрицы расстояний $D(N,N)$ получаются преобразованием $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$.

Пусть теперь задана матрица расстояний $D(N,N)$. Чтобы получить близости, необходимо назначить начало координат как некоторый объект ω_0, относительно которого можно по теореме косинусов вычислить скалярные произведения $s_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$, где для ненормированных величин диагональные элементы $s_{ii} = d_{0i}^2$ представляют расстояния всех остальных объектов $\omega_i \in \Omega,\ \ i = 1, \ldots N$ до начала координат ω_0.

Начало координат ω_0 можно выбрать разными способами, например, по методу главных проекций Торгерсона [3] поместить его в центр тяжести множества Ω. Тогда объект ω_0 будет представлен своими расстояниями до остальных объектов следующим образом:

$$d_{0i}^2 = d^2(\omega_0,\ \omega_i) = \frac{1}{N}\sum_{p=1}^{N} d_{ip}^2 - \frac{1}{2N^2}\sum_{p=1}^{N}\sum_{q=1}^{N} d_{pq}^2;\ \omega_p,\ \omega_q \in \Omega\ . \qquad (3)$$

Тем не менее такое представление окажется неудобным, так как относительно такого начала координат скалярные произведения s_{ij} могут оказаться как положительными, так и отрицательными. Необходимо так назначить начало координат, чтобы все скалярные произведения объектов относительно него были бы неотрицательными $s_{ij} \geqslant 0$. Только в этом случае мы можем рассматривать полученные значения именно как близости между элементами множества в метрическом пространстве. Содержательно это означает, что относительно такого начала координат все объекты должны располагаться в положительном квадранте координатного пространства, т. е. такое новое начало координат должно располагаться вне выпуклой оболочки, образованной данным множеством объектов, и на некотором достаточном удалении от множества объектов.

Как и ранее, отметим, что второе слагаемое в (3) представляет собой дисперсию множества:

$$\sigma^2 = \frac{1}{2N^2} \sum_{p=1}^{N} \sum_{q=1}^{N} d_{pq}^2 \, .$$

Рассмотрим первое слагаемое в (3). Рассмотрим расстояния $d_{ip},\ p = 1, \ldots N$ от объекта ω_i до остальных объектов ω_p как компоненты соответствующего вектора в некотором N-мерном пространстве, которое удобно считать «вторичным». Тогда величина $\sum_{p=1}^{N} d_{ip}^2$ представляет собой квадрат нормы этого вектора, т.е. квадрат расстояния от начала координат, а первое слагаемое из (3) представляет собой среднее квадрата этой нормы.

Таким образом, начало координат в таком вторичном пространстве будет представлено как объект ω_{0^0} своими расстояниями до остальных объектов $d_{0^0 i}^2,\ i = 1, \ldots N$, инвариантными относительно размера множества. Тогда начало координат по методу Торгерсона будет представлено как объект ω_0 своими расстояниями $d_{0i}^2 = d_{0^0 i}^2 - \Delta$, где $\Delta = \sigma^2$.

Легко увидеть, что изменение константы Δ позволит получить начало координат не в центре тяжести множества. При $\Delta = 0$ начало координат как объект ω_0 будет максимально удалено от центра тяжести множества. Условие $\Delta = 0$ можно понимать как наименьший разброс элементов множества по сравнению с расстояниями до начала координат. Очевидно, что в этом случае скалярные произведения будут близки к единице, если получившиеся расстояния до начала координат окажутся значительными. Если $\Delta < 0$, то это свойство только усилится.

При $\Delta > \sigma^2$ обязательно возникнет ситуация, когда не удастся получить корректный вид матрицы скалярных произведений $S(N, N)$. Это произойдет, когда некоторые из расстояний $d_{0i}^2 = d_{0^0 i}^2 - \Delta$ до начала координат окажутся нулевыми или отрицательными. В этом случае условие $\Delta > \sigma^2$ можно понимать как увеличенный по сравнению с реальным разброс элементов множества. В итоге допустимые значения величины Δ находятся в интервале $0 \leqslant \Delta \leqslant \sigma^2$.

Эксперименты

Были проведены эксперименты на данных по ирисам [10], которые представляют собой измерения четырех признаков (длина и ширина чашелистика, длина и ширина лепестка) пятидесяти экземпляров растений каждого из трех видов (Iris Setosa — «касатик щетиноносный», Iris Versicolor — «касатик разноцветный», Iris Virginica — «касатик виргинский»), всего 150 экземпляров.

Известно, что первый класс (Iris Setosa) хорошо отделен от остальных двух классов (второй класс — Iris Versicolor, третий класс — Iris Virginica), которые слегка пересекаются между собой. Так как классификация объектов для этих данных заранее известна,

Таблица 1. Ирисы. Разделение классов

Нач. разбиение	Ошибки ($\alpha = 0$)	α_{opt}	Ошибки (α_{opt})
50-50-50	16	3–6	15
50-70-30	16	3–6	15
50-30-70	16	3–6	15
50-50	16	12–17.7	15
70-30	16	12–17.7	15
30-70	16	12–17,7, 22–22,4	15

то необходимо показать, что применение двухкомпонентной функции критерия качества кластеризации позволяет объективно улучшить результат разбиения, правильно отделив первый класс от остальных и уменьшив ошибки кластеризации для второго и третьего классов.

Для примера рассмотрим критерий $\tilde{J}_\delta(K) = \tilde{J}(K) + \alpha\delta(K)$, где нужно будет подобрать масштабирующий коэффициент α для дополнительной части.

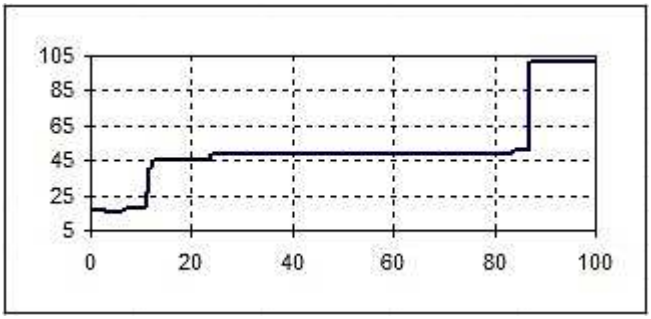
Известно, что алгоритм k-средних дает локально-оптимальный результат, который зависит от начального разбиения. Поэтому в экспериментах начальные разбиения фиксировались заранее, чтобы иметь возможность для сравнения результатов. Для трех классов использовались начальные разбиения: 50-50-50 (исходное), 50-70-30 (20 объектов третьего класса ошибочно отнесены ко второму классу) и 50-30-70 (20 объектов второго класса ошибочно отнесены к третьему классу). Для двух пересекающихся классов (второй и третий) использовались начальные разбиения: 50-50 (исходное), 70-30 (20 объектов третьего класса ошибочно отнесены ко второму классу) и 30-70 (20 объектов второго класса ошибочно отнесены к третьему классу).

В первой серии экспериментов было показано (см. табл. 1 и рис. 1), что для всех начальных разбиений удается подобрать оптимальное значение коэффициента α, при котором для случая трех классов первый отделяется безошибочно, а число ошибок разделения второго и третьего классов уменьшается по сравнению с классическим критерием ($\alpha = 0$). То же самое было показано и для случая двух пересекающихся классов (без первого). В обоих экспериментах ошибочными оказались одни и те же объекты: 102, 107, 114, 115, 120, 122, 124, 127, 128, 134, 135, 139, 143, 147, 150. Также видно, что оптимальное значение коэффициента α зависит от соотношения дисперсий разделяемых множеств.

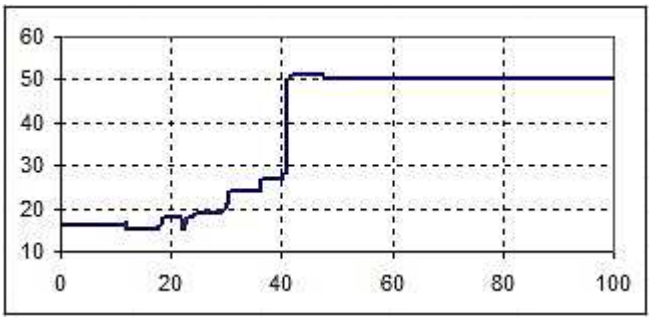
Во второй серии экспериментов рассматривалась известная проблема, когда разделение резко различающихся по размеру кластеров приводит к ошибкам, так как критерий $\tilde{J}(K)$ стремится разбить большой кластер, образованный вторым и третьем классами, и увеличить размер небольшого кластера, образованного первым классом. Таким образом, при $\alpha = 0$ три объекта 58, 94 и 99 были неправильно отнесены к первому классу.

Здесь также удается подобрать оптимальное значение коэффициента $\alpha = 27$, при котором достигается безошибочное разделение. В данном случае небольшой первый класс безошибочно отделился от второго и третьего классов, рассмотренных вместе как большой класс (рис. 1).

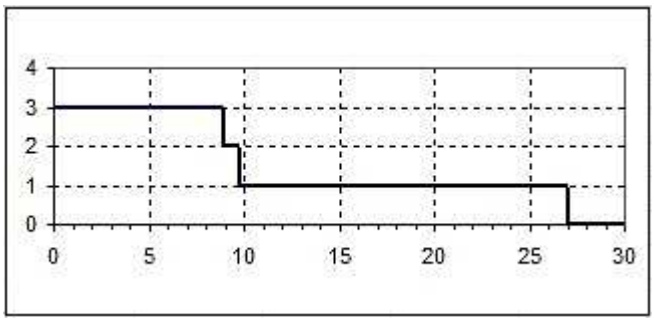
В случае, когда отсутствует признаковое пространство, множество элементов представлено только результатами их парных сравнений в виде матрицы близостей или расстояний.



Setosa – Versicolor – Virginica (50-50-50, 50-70-30, 50-30-70)



Versicolor – Virginica (30-70)



Setosa – Versicolor&Virginica (50-100, 100-50, 30-120)

Рис. 1. Ирисы. Графики числа ошибок при подборе оптимального значения α

В данной работе рассмотрены новые версии известного алгоритма k-средних для случая, когда пространство исходных признаков неизвестно. Иногда оказывается, что в этом случае удобнее не вводить понятие среднего объекта, так как возможны трудности с его интерпретацией, а применять перестановочные версии агрегирующих процедур.

В данной работе построены корректные версии таких процедур, которые дают одинаковый результат, если полученное разбиение окажется все-таки помещенным в подходящее признаковое пространство.

Применение двухкомпонентной целевой функции позволяет улучшить качество кластеризации, что было показано на известных данных по ирисам.

Очевидно, что в таких перестановочных алгоритмах необходимо снижать их сложность, выполняя пересчет значений критерия на основе приращений при переносах объектов между кластерами. Это можно делать экономно на основе рекуррентных соотношений, и такие способы известны.

Литература

[1] *Duda R. O., Hart P. E., Stork D. G.* Pattern classification. N.Y.: Wiley, 2001. 654 p.

[2] *Шлезингер М. И.* О самопроизвольном различении образов // *Читающие автоматы и распознавание образов.* Киев: Наукова думка, 1965. С. 38–45.

[3] *Torgerson W. S.* Theory and methods of scaling. N.Y.: Wiley, 1958. 460 p.

[4] *Двоенко С. Д.* Кластеризация элементов множества на основе взаимных расстояний и близостей // *ММРО-13.* М: МАКС-Пресс, 2007. С. 114–117.

[5] *Двоенко С. Д.* Кластеризация множества, описанного парными близостями и расстояниями между его элементами // *Сибирский журнал индустриальной математики*, 2009. Т. 12, № 1. С. 61–73.

[6] *Dvoenko S. D.* Clustering and separating of a set of members in terms of mutual distances and similarities // *Trans. Machine Learning Data Mining*, 2009. Vol. 2, No. 2. P. 80–99.

[7] *Dvoenko S. D.* On featureless k-means clustering // *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*, 2013. Tilburg, the Netherlands. P. 70.

[8] *Owsinski J. W.* The bi-partial approach in clustering and ordering: the model and the algorithms // *Statistica & Applicazioni. Special Issue*, 2011. P. 43–59.

[9] *Owsinski J. W.* The matter of scale: Perceiving distances and proximities in the bi-partial clustering setting // *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*, 2013. Tilburg, the Nethelands. P. 88–89.

[10] *Fisher R. A.* The use of multiple measurements in taxonomic problems // *Ann. Eugenics*, 1936. Vol. 7, No. 9. P. 179–188.

References

[1] *Duda R. O., Hart P. E., Stork D. G.* 2001. Pattern classification. N.Y.: Wiley. 654 p.

[2] *Schlesinger M. I.* 1965. On spontaneous pattern distinguishing. *Reading Automata and Pattern Recognition.* Kiev: Naukova Dumka. 38–45. (in Russ.)

[3] *Torgerson W. S.* 1958. Theory and methods of scaling. N.Y.: Wiley. 460 p.

[4] *Dvoenko S. D.* 2007. Clustering a set members by mutual distances ans similarities. *MMPR-13.* M: MAKS-Press. 114–117. (in Russ.)

[5] *Dvoenko S. D.* 2009. Clusterization of the set presented by distances and similarities between its elements. *Syberian Journal of Industrial Mathematics* 12(1):61–73. (in Russ.)

[6] *Dvoenko S. D.* 2009. Clustering and separating of a set of members in terms of mutual distances and similarities. *Trans. Machine Learning Data Mining* 2(2):80–99.

[7] *Dvoenko S. D.* 2013. On featureless k-means clustering. *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*. Tilburg, the Nethelands. 70.

[8] *Owsinski J. W.* 2011. The bi-partial approach in clustering and ordering: the model and the algorithms. *Statistica & Applicazioni. Special Issue*. 43–59.

[9] *Owsinski J. W.* 2013. The matter of scale: Perceiving distances and proximities in the bi-partial clustering setting. *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*. Tilburg, the Netherlands. 88–89.

[10] *Fisher R. A.* 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7(9):179–188.

Моделирование вариаций космических лучей и выделение аномалий на основе совмещения вейвлет-преобразования с нейронными сетями*

***О. В. Мандрикова*[1,2], *Т. Л. Заляев*[1]**
oksanam1@mail.ru;tim.aka.geralt@mail.ru
[1]Институт космофизических исследований и распространения радиоволн ДВО РАН, Паратунка, Камчатский край, Российская Федерация; [2]Камчатский государственный технический университет, г. Петропавловск-Камчатский, Российская Федерация

В работе исследованы данные нейтронных мониторов станций «Афины», «Новосибирск» и «Апатиты» за 2005–2013 гг. и обнаружены аномальные особенности, возникающие в вариациях космических лучей во время сильных магнитных бурь. Исследования основаны на разработанном авторами методе моделирования компонент данных космических лучей путем совмещения вейвлет-преобразования и нейронных сетей прямого распространения. Выполняется кратномасштабное вейвлет-разложение данных и выделяются информативные компоненты. Полученные компоненты аппроксимируются нейронными сетями прямого распространения. Метод позволяет выполнить детальный анализ структуры данных и путем анализа ошибок нейронной сети выявить аномальные особенности (Форбуш-эффекты) во временном ходе космических лучей.

Ключевые слова: *космические лучи; вейвлет-преобразование; нейронные сети; магнитные бури; Форбуш-эффекты*

Modeling of of cosmic ray variations and allocation of anomalies based on a combination of wavelet transform with neural networks*

***O. V. Mandrikova*[1,2], *T. L. Zalyaev*[1]**
[1]Institute of Cosmophysical Researches and Radio Wave Propagation, Far Eastern Branch of the Russian Academy of Sciences, Paratunka, Kamchatka Region, Russia; [2]Kamchatka State Technical University, Petropavlovsk-Kamchatsky, Russia

Background: Valuable information about the topology change of the geomagnetic field during magnetic storms is provided by study of the dynamics of cosmic rays. Observed on the Earth's surface variations of cosmic rays are the integral result of various solar, heliospheric and atmospheric phenomena and have a complex internal structure. The most significant changes in the parameters of cosmic rays are caused by coronal mass ejections and the following changes in the parameters of the interplanetary field and the solar wind.

In disturbed periods, the recorded parameters of the environment have a complex nonstationary structure, contain nonsmooth local features which occur at random time moments, and carry important information about the studied processes. Lack of theoretical apparatus providing an adequate description of the analyzed data leads to an inevitable loss and distortion of the information and requires advanced methods, among which are of great importance the methods of pattern recognition and digital signal processing.

*Работа выполнена при поддержке гранта РНФ №14-11-00194.

Methods: Based on a combination of multiresolution wavelet decompositions with neural network, a method of approximation of the cosmic rays time course and the allocation of anomalous variations (Forbush effects) that occur during the periods of high solar activity is proposed. The method allows to study in detail the structure of the data, to allocate informative components, and to build their approximation based on neural network.
Results: On the basis of the proposed method for the stations "Novosibirsk", "Apatity", and "Athens software systems for neural network approximation of typical variations of cosmic rays were built and the analysis of data in the periods of strong magnetic storms was made. Application of the method allowed to study the dynamic characteristics of the processes and to allocate anomalous effects related to solar activity.
Concluding Remarks: Application of the method in conjunction with other methods and approaches allows to better perform the assessment of the state of space weather.
Keywords: *cosmic rays; wavelet transform; neural networks; magnetic storms; Forbush effects*

Введение

Работа направлена на создание методов и алгоритмов анализа регистрируемых геофизических параметров и изучение процессов в околоземном пространстве. Сложный характер изучаемых процессов, их априорная неопределенность и, как следствие, сложная структура регистрируемых данных требует наличия целого комплекса методов и технологий, позволяющих выполнять моделирование, структурный анализ данных и интерпретацию получаемых результатов. Отсутствие теоретического аппарата, обеспечивающего адекватное описание анализируемых данных, приводит к неизбежной потере и искажению информации и требует применения современных методов, среди которых важное значение имеют методы распознавания образов и цифровой обработки сигналов [1, 2, 3, 4, 5, 6]. Исследования динамики потока космических лучей, являющихся предметом исследований в данной работе, позволяют получать ценную информацию об изменении топологии геомагнитного поля во время магнитных бурь. Наблюдаемые на поверхности Земли вариации космических лучей являются интегральным результатом различных солнечных, гелиосферных и атмосферных явлений и имеют сложную внутреннюю структуру. Наиболее существенные изменения в параметрах космических лучей вызывают выбросы коронарной массы и следующие за ними изменения в параметрах межпланетного поля и солнечного ветра [7]. Для изучения их динамических свойств в настоящее время получают развитие методы адаптивной аппроксимации, вейвлет-преобразование и нейронные сети. Использование нейронных сетей при первичной обработке данных нейтронных мониторов позволило повысить эффективность процедуры подавления шума, по сравнению с медианными методами [6]. На основе совмещения вейвлет-преобразования с методом разложения на эмпирические моды в долгосрочных временных изменениях хода космических лучей выделены доминирующие временные масштабы (периоды 11 лет, 22 года, 6 лет и двухлетние колебания) и определена их физическая природа [7]. В данной работе на основе совмещения кратномасштабных вейвлет-разложений с нейронными сетями построены аппроксимации компоненты временного хода космических лучей для различных станций регистрации данных. Описаны этапы выделения информативных компонент данных и построения нейронной сети. Выполненный анализ временного хода данных космических лучей в периоды сильных магнитных бурь (анализировались магнитные бури 5–7 апреля 2010 г. и 17 марта 2013 г.) показал, что возникающие в вариациях космических лучей аномальные изменения формируются на фоне повышенной геомагнитной

активности. В моменты существенного возрастания скорости солнечного ветра на анализируемых станциях выделены локальные возрастания уровня космических лучей (предповышения) и интенсивности геомагнитных возмущений. Отмечено, что регистрируемые на различных станциях данные космических лучей, как правило, имеют общий характер изменения. В анализе использовались минутные данные международной сети магнитных обсерваторий INTERMAGNET (`www.intermagnet.org`) и данные нейтронных мониторов, полученные в рамках проекта NMDB (`www.nmdb.eu/`).

Выделение компонент данных и построение нейронных сетей

Кратномасштабные разложения данных на компоненты. На основе кратномасштабных вейвлет-разложений до уровня получаем представление данных в виде [8, 9, 10]:

$$f_0(t) = \sum_{j=-1}^{-m} f^d[2^j t] + f^a[2^{-m} t]\,, \tag{1}$$

где исходное разрешение данных $j = 0$, $f^d[2^j t] \in W_j$, $f^a[2^{-m}t] \in V_{-m}$, $W_j = \mathrm{close}_{L^2(R)}(2^{j/2}\Psi(2^j t - n)) : n \in Z)$; Ψ — базисный вейвлет; $V_{-m} = \mathrm{close}_{L^2(R)}(2^{j/2}\varphi(2^j t - n)) : n \in Z)$; φ — скэйлинг-функция; j — разрешение. Компонента ряда $f^a[2^{-m}t] = \sum_n c_{-m,n}\varphi_{-m,n}$, где $c_{-m,n} = \langle f, \varphi_{-m,n}(t)\rangle$, является сглаженной компонентой, компоненты $f^d[2^j t] = \sum_n d_{j,n}\Psi_{j,n}(t)$, где $d_{j,n} = \langle f, \Psi_{j,n}\rangle$, являются разномасштабными детализирующими компонентами.

После восстановления исходного разрешения $j = 0$, полученные после преобразования (1) компоненты имеют представление:

$$f_0^{a,-m} = \sum_n c_{0,n}^{-m}\varphi_{0,n}(t),\ f_0^{d,j}(t) = \sum_n d_{0,n}^{j}\Psi_{0,n}(t)\,,$$

где верхние индексы $(-m), j$ соответствуют разрешению компоненты до выполнения операции вейвлет-восстановления.

Аппроксимация сглаженной компоненты данных нейронной сетью. Для сглаженной компоненты $f_0^{a,-m}$, используя *сеть переменной структуры* (*сеть переменной структуры* — это многослойная прямонаправленная сеть, архитектура которой определяется путем минимизации ошибки решения на множестве обучающих векторов [2]), строим отображение

$$y : f_0^{a,-m} \to f_0^{*a,-m}\,.$$

Если в построенном отображении $\hat{f}_0^{*(-m)}$ действительный выход сети, а $f_0^{*(-m)}$ — желаемый, то $f_0^{*(-m)} = y(f_0^{(-m)})$ — неизвестная функция, а $\hat{f}_0^{*(-m)}$ — ее аппроксимация, которую воспроизводит нейронная сеть. При подаче на вход обученной нейронной сети значений функции $f_0^{(-m)}$ из интервала $[t_n - Q + 1, t_n]$, сеть становится способной вычислить упрежденные ее значения на временном интервале $[t_n + 1, t_n + I]$, где t_n — текущий дискретный момент времени; I — длина интервала упреждения.

Ошибка сети (ошибка аппроксимации) в момент времени t_n определяется как разность между желаемым $f_0^{*(-m)}$ и действительным $\hat{f}_0^{*(-m)}$ выходными значениями функции.

$$e_m[t_n] = \sum_{i=1}^{I} f_{0,i}^{*(-m)}[t_n] - \hat{f}_{0,i}^{*(-m)}[t_n]\,,$$

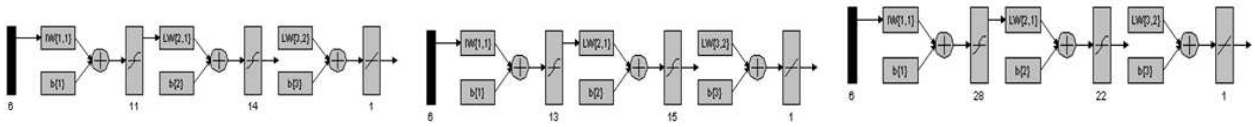


Рис. 1. Архитектура нейронных сетей по аппроксимации данных КЛ для станций «Афины», «Апатиты» и «Новосибирск»

где i — шаг упреждения данных, квадратные скобки обозначают дискретные моменты времени. Разработанный авторами алгоритм построения сети и выбора уровня вейвлет-разложения, основанный на минимизации ошибки аппроксимации, приведен в работе [11].

Если во временном ходе данных возникает *аномальное изменение*, то абсолютное значение ошибки сети возрастет. Поэтому *выделение аномальных изменений* может быть основано, например, на проверке условия:

$$E_{m,U} = \frac{1}{U}\sum_{n=1}^{U} e_m[t_n] > T\,.$$

Моделирование данных космических лучей

В экспериментах использовались минутные данные космических лучей за период 2005–2013 гг. Следуя критериям выбора аппроксимирующих вейвлетов, предложенным в работе [9], для кратномасштабных вейвлет-разложений использовались вейвлеты семейства Койфлеты. В частности, в работе [11] показано, что при совместном применении вейвлет-преобразования и нейронных сетей, наименьшую погрешность аппроксимации вариаций космических лучей, позволяют получить Койфлеты порядка 3. Поскольку динамика космических лучей существенно зависит от электромагнитной обстановки в солнечной системе и находит отражение в геомагнитном поле [12], обучающие множества нейронных сетей формировались из данных, регистрируемых в периоды спокойного геомагнитного поля. Обучение сетей выполнялось на основе алгоритма обратного распространения ошибки [13].

Архитектура построенных нейронных сетей для станций «Афины», «Новосибирск» и «Апатиты» представлена на рис. 1. Как следует из рис.1, построенные нейронные сети имеют трехслойную структуру и выполняют следующее преобразование данных:

$$c_{j,n+1}(t) = \alpha^3_\chi\left(\sum_s \omega^3_{\chi s}\alpha^2_s\left(\sum_l \omega^2_{sl}\alpha^1_l\left(\sum_n \omega^1_{ln}c_{j,n}(t)\right)\right)\right),$$

где ω^1_{ln} — весовые коэффициенты нейрона входного слоя сети; ω^2_{sl} — весовые коэффициенты нейрона скрытого слоя сети; $\omega^3_{\chi s}$ — весовые коэффициенты нейрона выходного слоя; $\alpha^1_l(z) = \alpha^2_s(z) = (2/(1+\exp(-2z)) - 1$; $\alpha^3_\chi(z) = a * z + b$.

В табл. 1 показаны среднеквадратичные ошибки сетей, рассчитанные для спокойного (21.11.2013–23.11.2013) и возмущенного (16.03.2013–18.03.2013) периодов времени. На рис. 2–4 представлены результаты моделирования данных в эти периоды. Анализ результатов показывает, что в возмущенные периоды происходит изменение временного хода процесса, и ошибки сетей существенно увеличиваются.

Таблица 1. Среднеквадратичные ошибки нейронных сетей

Период времени	Афины	Новосибирск	Апатиты
16.03.2013–18.03.2013	0,0009	0,0008	0,0003
21.11.2013–23.11.2013	0,00005	0,00005	0,00004

Анализ данных в период сильных магнитных бурь

Первая анализируемая магнитная буря была зарегистрирована на Земле 5 апреля 2010 г. в 08.26 UT как внезапное начало SC магнитной бури. Скорость солнечного ветра возросла до 750–900 км/с. Примерно через полчаса в магнитосфере Земли возникла интенсивная суббуря, наблюдаемая в глобальном масштабе [14]. Как показывает анализ рис. 5 и 6, в эти моменты времени на анализируемых станциях возникли возрастания интенсивности геомагнитных возмущений (интенсивность геомагнитных возмущений оценивалась в соотвествии с методом предложенным в работе [15]) и локальные возрастания уровня космических лучей (ошибки нейронных сетей увеличились в 4 раза на станции «Афины» и в 2,5 раза на станции «Новосибирск», по сравнению со спокойным периодом). На станции «Афины» зафиксирован короткий Форбуш-эффект (понижение уровня космических лучей) с быстрым восстановлением, пик которого пришелся на 20:00 UT (наблюдается возрастание ошибки нейронной сети в 3 раза, по сравнению со спокойным периодом). На станции «Новосибирск» Форбуш-эффект был более длительным, пик его пришелся на 16:00 UT (возрастание ошибки нейронной сети в 10 раз, по сравнению со спокойным периодом).

На рис. 7 и 8 показаны результаты работы нейронных сетей для станций «Новосибирск» и «Афины» в период сильных магнитных бурь, произошедших 7–10 марта и 15–17 марта 2012 г. Форбуш-эффект в период бури 7–10 марта возник 8 марта и выразился в сильном понижении уровня космических лучей (до 10%) на обеих станциях. Во время Форбуш-эффекта существенно возрастают абсолютные значения ошибок сетей (до 10 раз), что позволяет автоматически фиксировать данные моменты времени. Сопоставление с геомагнитными данными показывает, что в периоды аномальных изменений хода космических лучей наблюдаются наиболее сильные геомагнитные возмущения. В период магнитной бури 15–17 марта моменты сильных возрастаний интенсивности геомагнитных возмущений совпадают с локальными понижениями уровня космических лучей и имеют более яркое проявление на станции «Новосибирск». Восстановление уровня космических лучей происходит после окончания бури.

Анализируемая на рис. 9 и 10 магнитная буря была зафиксирована на Земле 17 марта 2013 г. Скорость солнечного ветра достигла значений 700–750 км/с в связи с приходом ускоренного потока от CME 15 марта 2013 г. Приход ударной волны произошел 17.03.13 в 6:00 UT и зафиксирован сетью станции «Афины» как локальное изменение хода космических лучей (уровень ошибки нейронной сети в «Афинах» вырос до 20 раз, по сравнению со спокойным периодом). Одновременно с этими событиями существенно возросла интенсивность геомагнитных возмущений. Максимальные значения интенсивности возмущений поля зафиксированы 17 марта в период понижения уровня космических лучей с 16:00 до 20:00 UT (на обеих станциях наблюдалось возрастание ошибки нейронной сети). Резкий скачок скорости солнечного ветра 22 марта в период с 12:00 до 18:00 UT привел к повторному повышению уровня ошибок нейронных сетей (в 4 раза на станции «Новоси-

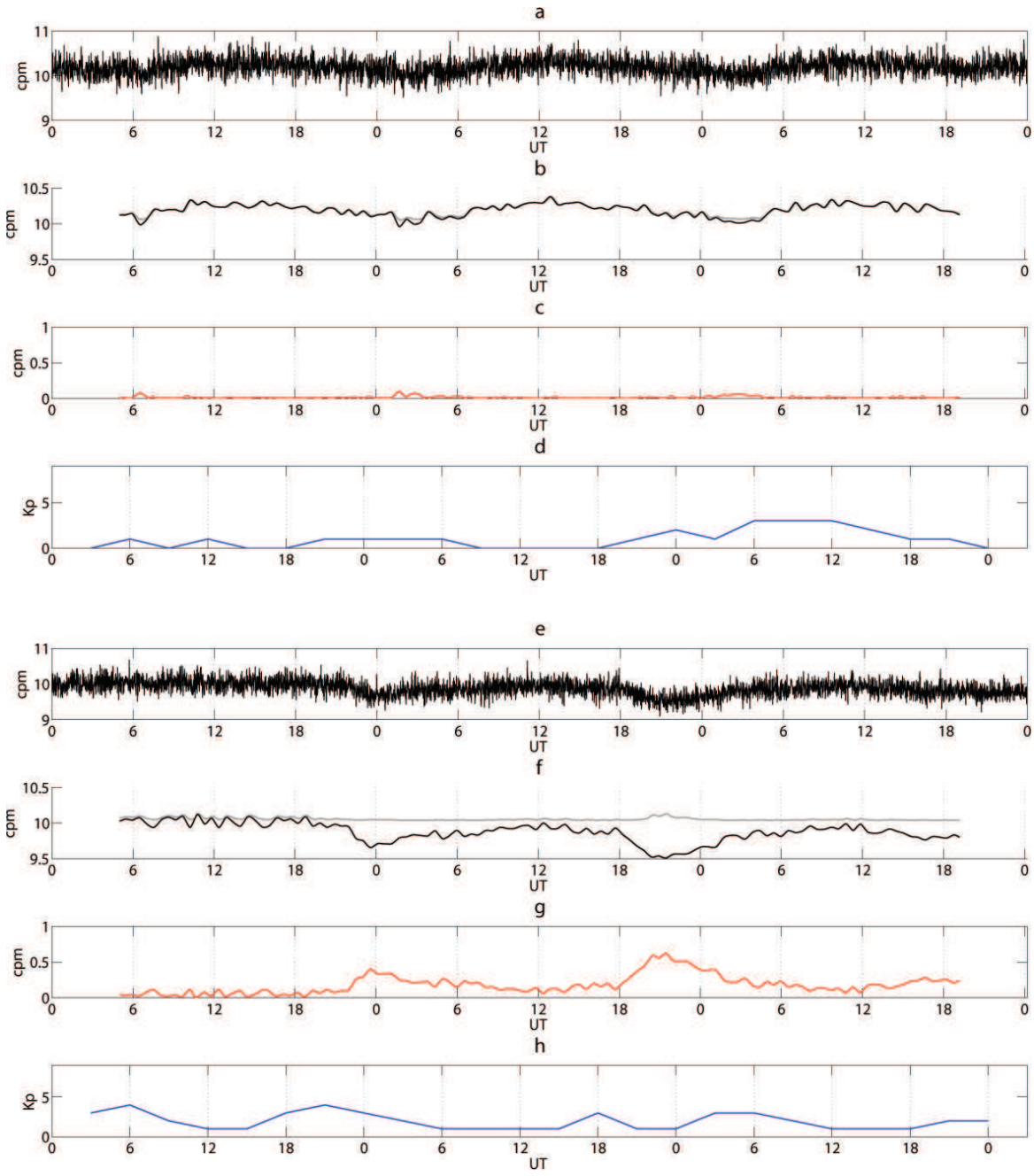


Рис. 2. Результаты работы нейронной сети, станция «Апатиты»: (a) сигнал космических лучей станции «Апатиты» за период 21–23 ноября 2013 г.; (b) сглаженная компонента вариации КЛ a (черный цвет) и ее аппроксимация нейронной сетью (серый цвет) (период спокойного геомагнитного поля); (c) абсолютные значения ошибок нейронной сети; (d) Kp индекс; (e) сигнал космических лучей станции «Апатиты» за период 14–16 декабря 2013 г.; (f) сглаженная компонента вариации КЛ a (черный цвет) и ее аппроксимация нейронной сетью (серый цвет) (период спокойного геомагнитного поля); (g) абсолютные значения ошибок нейронной сети; (h) Kp индекс

бирск», и в 40 раз на станции «Афины»). Возрастание интенсивности геомагнитного поля началось на несколько часов позже.

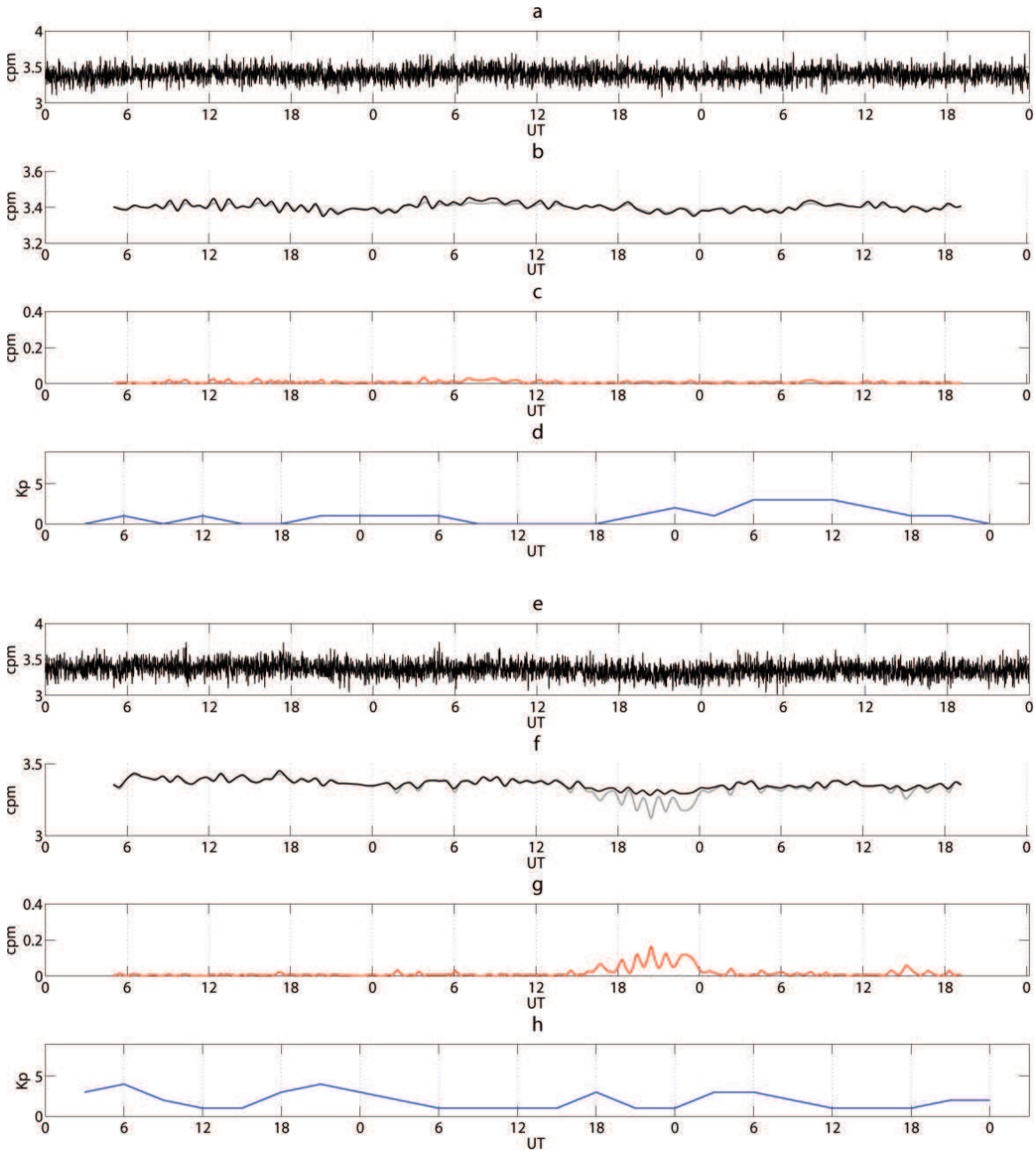


Рис. 3. Результаты работы нейронной сети, станция «Афины»: (a) сигнал космических лучей станции «Афины» за период 21–23 ноября 2013 г.; (b) сглаженная компонента вариации КЛ a (черный цвет) и ее аппроксимация нейронной сетью (серый цвет) (период спокойного геомагнитного поля); (c) абсолютные значения ошибок нейронной сети; (d) Kp индекс; (e) сигнал космических лучей станции «Апатиты» за период 14–16 декабря 2013 г.; (f) сглаженная компонента вариации КЛ a (черный цвет) и ее аппроксимация нейронной сетью (серый цвет) (период спокойного геомагнитного поля); (g) абсолютные значения ошибок нейронной сети; h)Kp индекс

Заключение

На основе разработанного авторами метода моделирования данных нейтронных мониторов построены нейросетевые программные системы по аппроксимации вариаций космических лучей для различных станций регистрации и выполненный анализ временного хода данных в периоды сильных магнитных бурь. Разработанный метод позволяет вы-

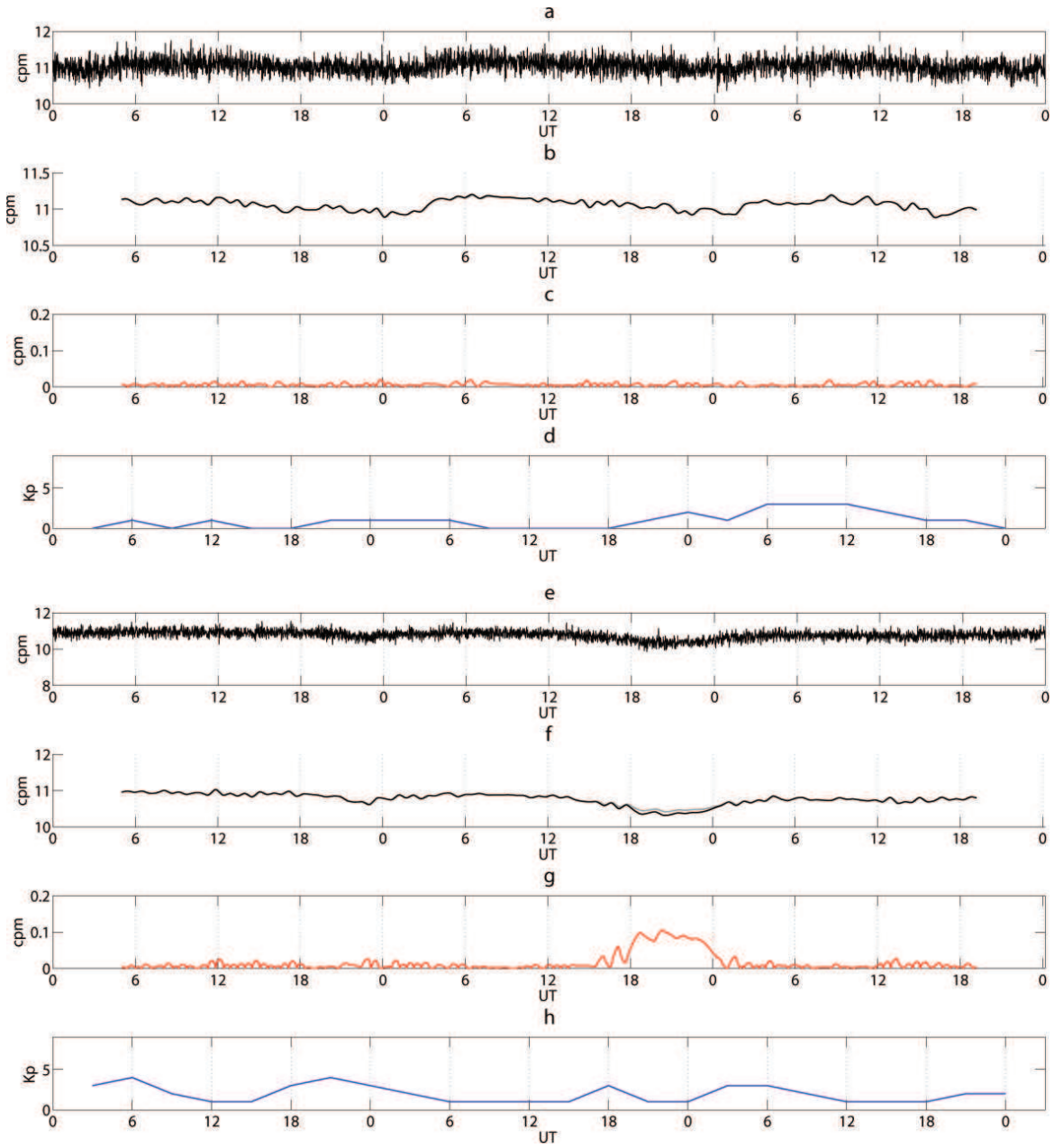


Рис. 4. Результаты работы нейронной сети, станция «Новосибирск»: (a) сигнал космических лучей станции «Новосибирск» за период 21–23 ноября 2013 г.; (b) сглаженная компонента вариации КЛ a (черный цвет) и ее аппроксимация нейронной сетью (серый цвет) (период спокойного геомагнитного поля); (c) абсолютные значения ошибок нейронной сети; (d) Kp индекс; (e) сигнал космических лучей станции «Апатиты» за период 14–16 декабря 2013 г.; (f) сглаженная компонента вариации КЛ a (черный цвет) и ее аппроксимация нейронной сетью (серый цвет) (период спокойного геомагнитного поля); (g) абсолютные значения ошибок нейронной сети; (h) Kp индекс

делять моменты возникновения Форбуш-эффектов и определять их продолжительность. Показано, что возникающие в вариациях космических лучей аномальные изменения формируются на фоне повышенной геомагнитной активности. Результаты обработки показали, что в моменты возрастания скорости солнечного ветра на анализируемых станциях фиксируются аномальные изменения во временном ходе космических лучей, что подтвер-

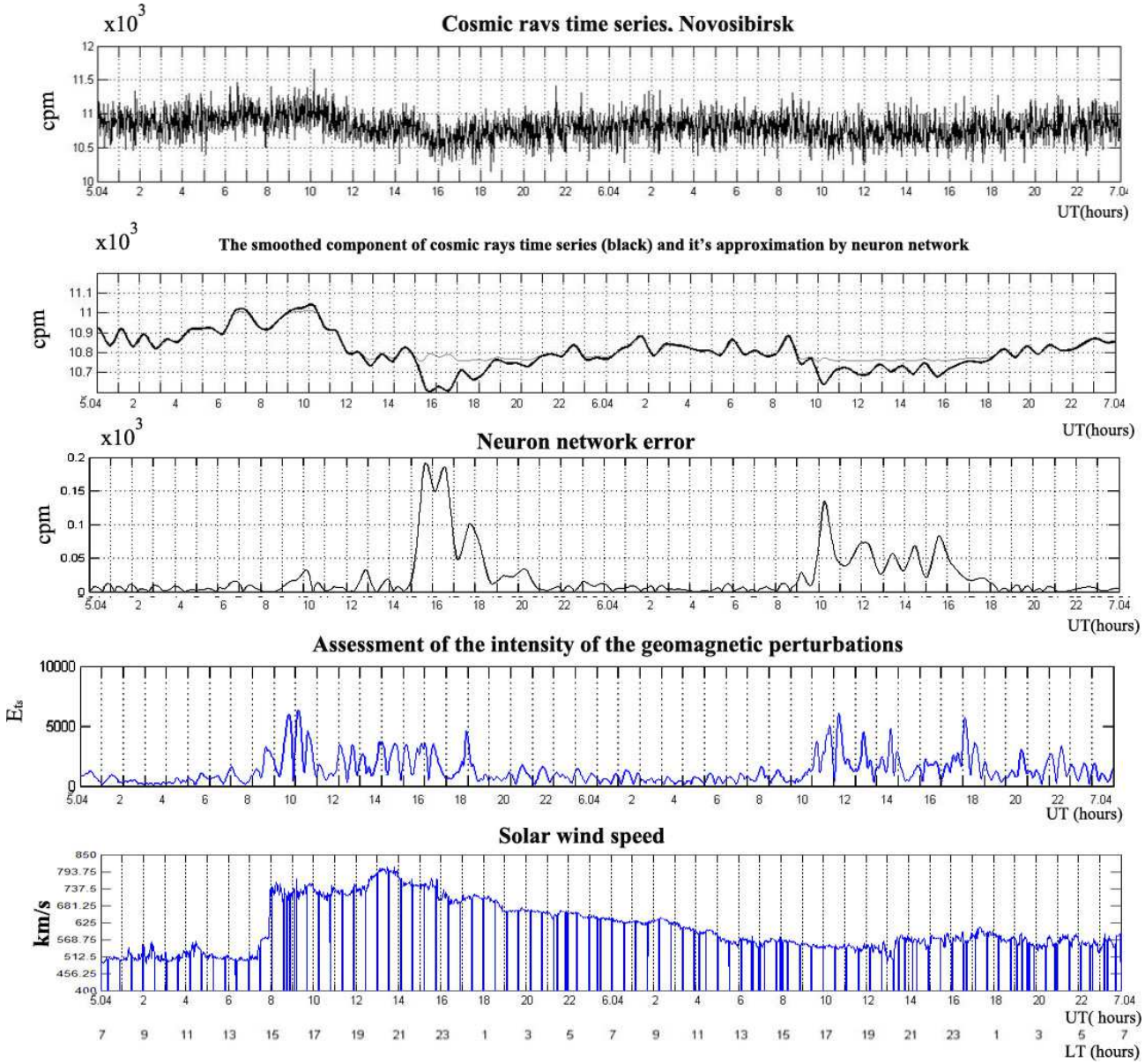


Рис. 5. Результаты анализа данных космических лучей станции «Новосибирск» за период с 05.04.2010 г. по 07.04.2010 г.

ждает существенное влияние скорости солнечного ветра на измеряемый на Земле уровень космических лучей.

Литература

[1] *Macpherson K. P., Conway A. J., Brown J. C.* Prediction of solar and geomagnetic activity data using neural networks // *J. Geophys. Res.*, 2001. Vol. 100. P. 735–744.

[2] *Nayar S. R. P., Radhika V. N., Seen P. T.* Seena Investigation of substorms during geomagnetic storms using wavelet Techniques // *ILWS Workshop Proceedings*, Goa, India, 2006.

[3] *Jach A., Kokoszka P., Sojka J., Zhu L.* Wavelet-based index of magnetic storm activity // *J. Geophys. Res.*, 2006. 111. doi:10.1029/2006ja011635. `http://onlinelibrary.wiley.com/doi/10.1029/2006JA011635/pdf`.

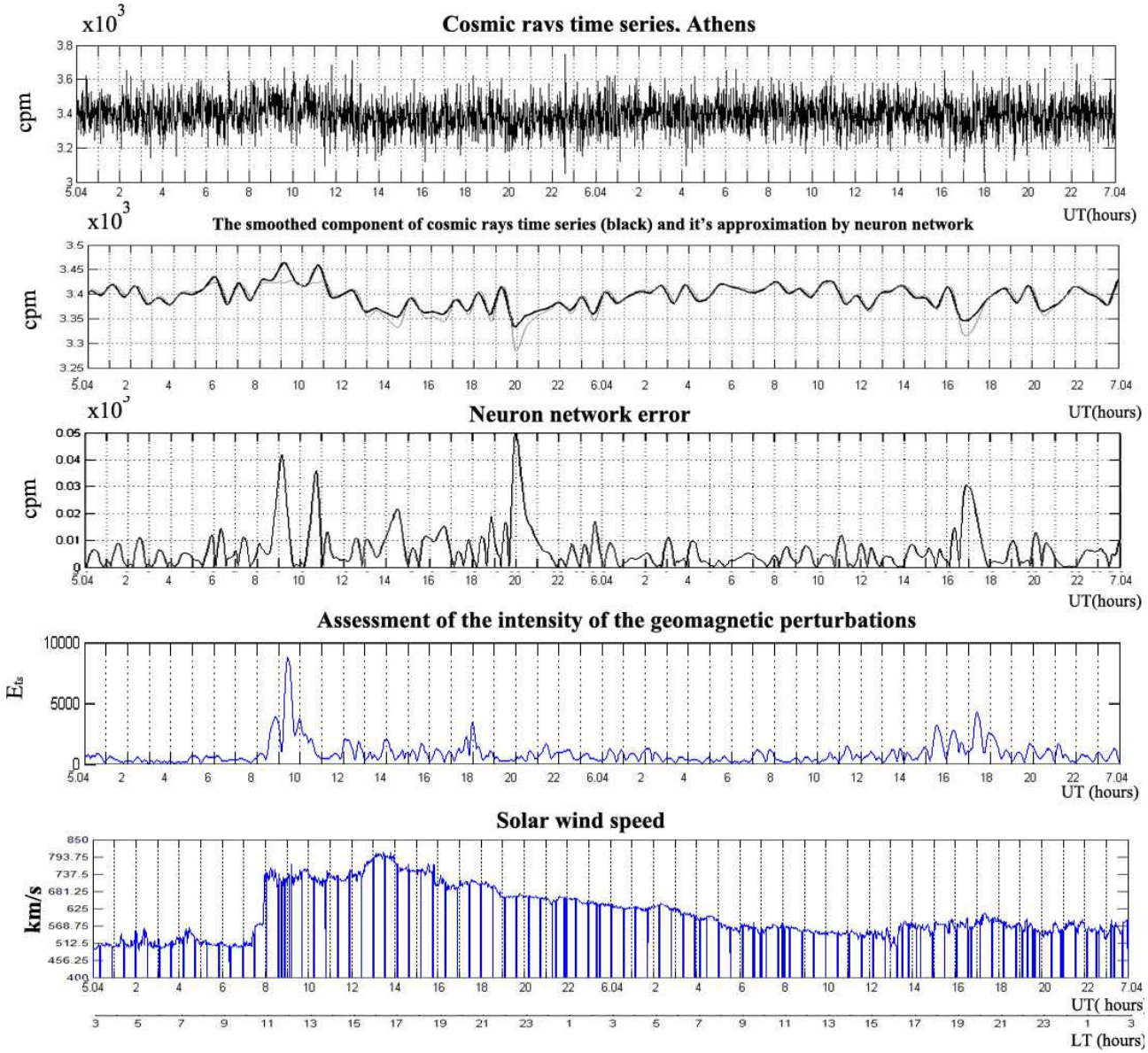


Рис. 6. Результаты анализа данных космических лучей станции «Афины» за период с 05.04.2010 г. по 07.04.2010 г.

[4] *Hafez A. G., Ghamry E., Yayama H., Yumoto K.* Systematic examination of the geomagnetic storm sudden commencement using multi resolution analysis // *Advances in Space Research*, 2013. Vol. 51. P. 39–49.

[5] *Xu Z., Zhu L., Sojka J., Kokoszka P., Jach A.* An assessment study of the wavelet-based index of magnetic storm activity (WISA) and its comparison to the Dst index // *J. Atmos. Solar-Terr. Phys.*, 2008. Vol. 70. P. 1579-1588.

[6] *Paschalis P., Sarlanis C., Mavromichalaki H.* Artificial neural network approach of cosmic ray primary data processing // *Solar Physics*, 2013. Vol. 182, No. 1. P. 303–318.

[7] *Vecchio A., Laurenza M., Storini M., Carbone V* New insights on cosmic ray modulation through a joint use of non stationary data-processing methods // *J. Advances Astronomy*, 2012. doi:10.1155/2012/834247.

[8] *Chui C. K.* An introduction in wavelets. New York: Academic Press, 1992.

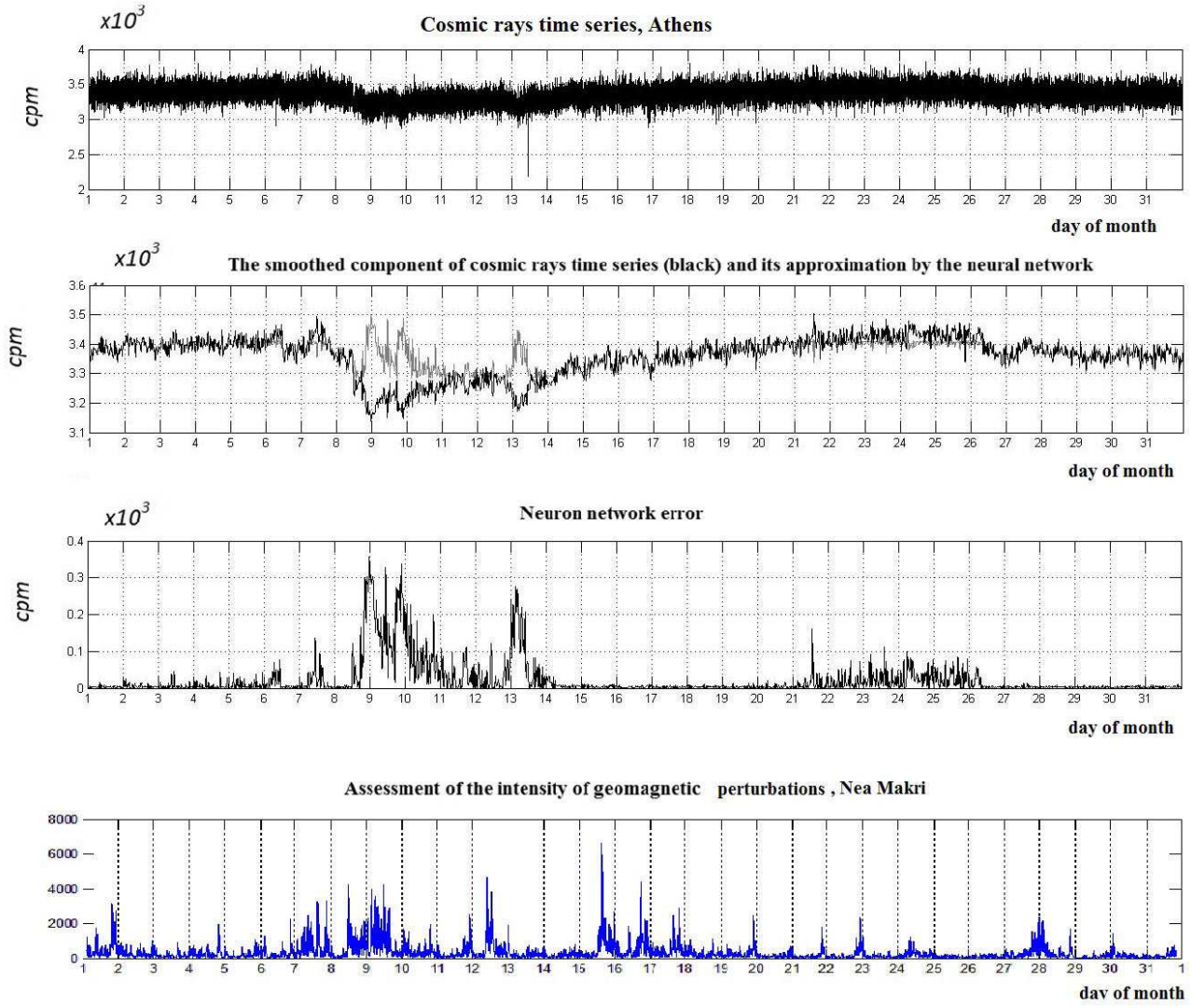


Рис. 7. Результаты работы нейронной сети на станции «Афины» за март 2012 г.

[9] *Daubechies I.* Ten lectures on wavelets. CBMS-NSF lecture notes ser. Philadelphia: SIAM, 1992. 61.

[10] *Mallat S.* A wavelet tour of signal processing. London: Academic Press, 1999.

[11] *Мандрикова О. В., Заляев Т. Л.* Моделирование вариаций космических лучей на основе совмещения кратномасштабного анализа и сетей переменной структуры // *Сб. тезисов докладов VI Междунар. научн.-технич. конф. по мягким вычислениям и измерениям (SCM'2013).* СПб, 2013. С. 111–117.

[12] *Akasofu S. I., Chapman S.* . Solar-terrestrial physics. Oxford: Oxford University Press, 1972.

[13] *Haykin S.* Neural Networks: A comprehensive foundation. 2nd ed. New York:Prentice-Hall, 1999.

[14] *Клейменова Н. Г. , Зелинский Н. Р. , Козырева О. В., Малышева Л. М., Соловъев А. А. , Богоутдинов Ш. Р.* Геомагнитные пульсации Pc3 на приэкваториальных широтах в начальную фазу магнитной бури 5 апреля 2010 г. // *Геомагнетизм и аэрономия*, 2013. Т. 53. С. 313–320.

[15] *Mandrikova O., Solovjev I., Geppenerc V., Taha Al-Kasasbehd R., Klionskiy D.* Analysis of the Earth's magnetic field variations on the basis of a wavelet-based approach // *Digital Signal Processing*, 2013. Vol. 23, No. I.1. P. 329—339.

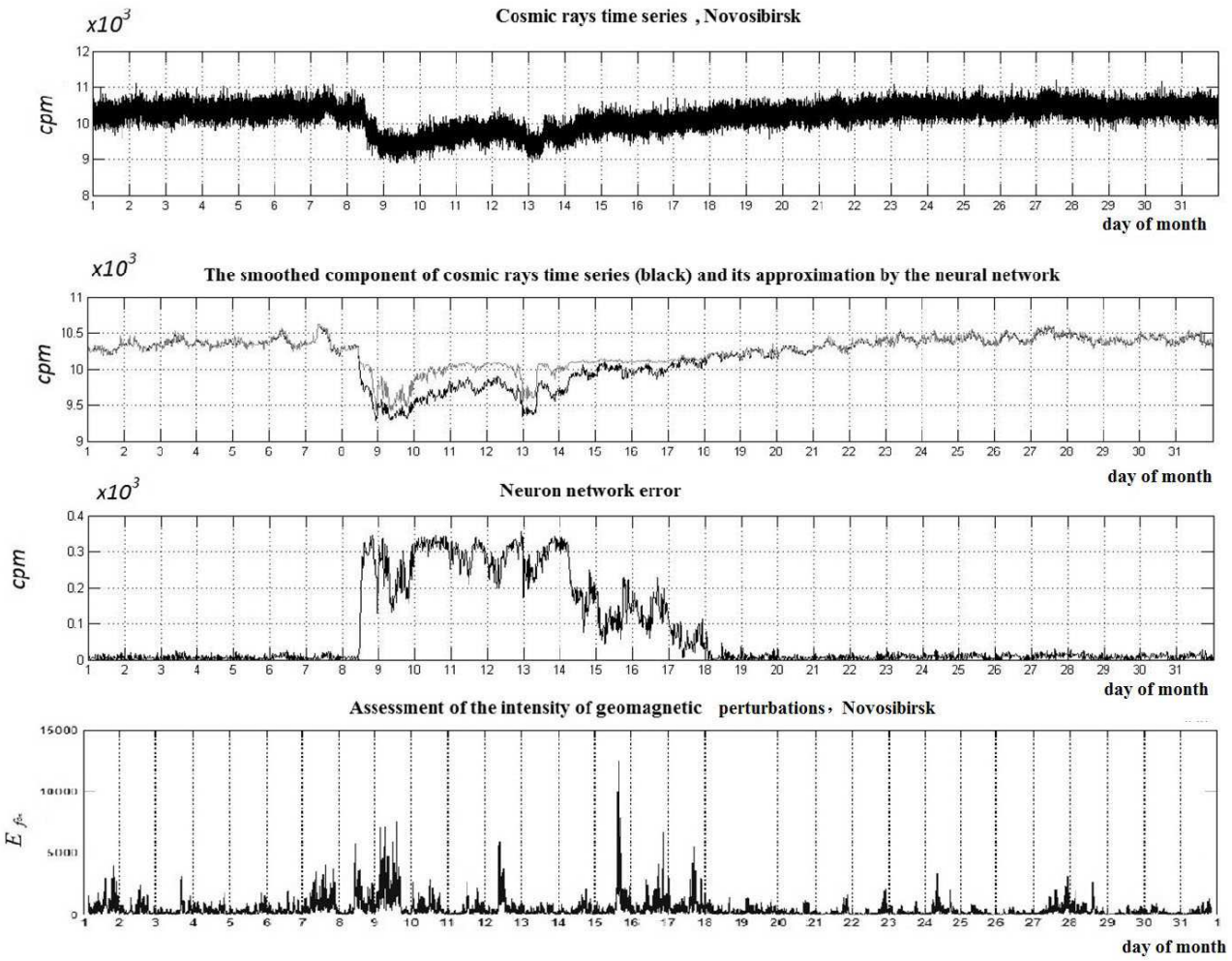


Рис. 8. Результаты работы нейронной сети на станции «Новосибирск» за март 2012 г.

References

[1] *Macpherson K. P., Conway A. J., and Brown J. C.* 2001. Prediction of solar and geomagnetic activity data using neural networks. *J. Geophys. Res.* 100:735–744.

[2] *Nayar S. R. P., Radhika V. N. and Seen P. T.* 2006. Seena investigation of substorms during geomagnetic storms using wavelet Techniques. *ILWS Workshop Proceedings*. Goa, India.

[3] *Jach A., Kokoszka P., Sojka J., Zhu L.* 2006. Wavelet-based index of magnetic storm activity. *J. Geophys. Res.* 111. doi:10.1029/2006ja011635. Available at: `http://onlinelibrary.wiley.com/doi/10.1029/2006JA011635/pdf`.

[4] *Hafez A. G., Ghamry E., Yayama H., Yumoto K.* 2013. Systematic examination of the geomagnetic storm sudden commencement using multi resolution analysis. *Advances Space Research* 51:39–49.

[5] *Xu Z., Zhu L., Sojka J., Kokoszka P., Jach A.* 2008. An assessment study of the wavelet-based index of magnetic storm activity (WISA) and its comparison to the Dst index. *J. Atmos. Solar-Terr. Phys.* 70:1579–1588.

[6] *Paschalis P., Sarlanis C., Mavromichalaki H.* 2013. Artificial neural network approach of cosmic ray primary data processing. *Solar Phys.* 182(1):303–318.

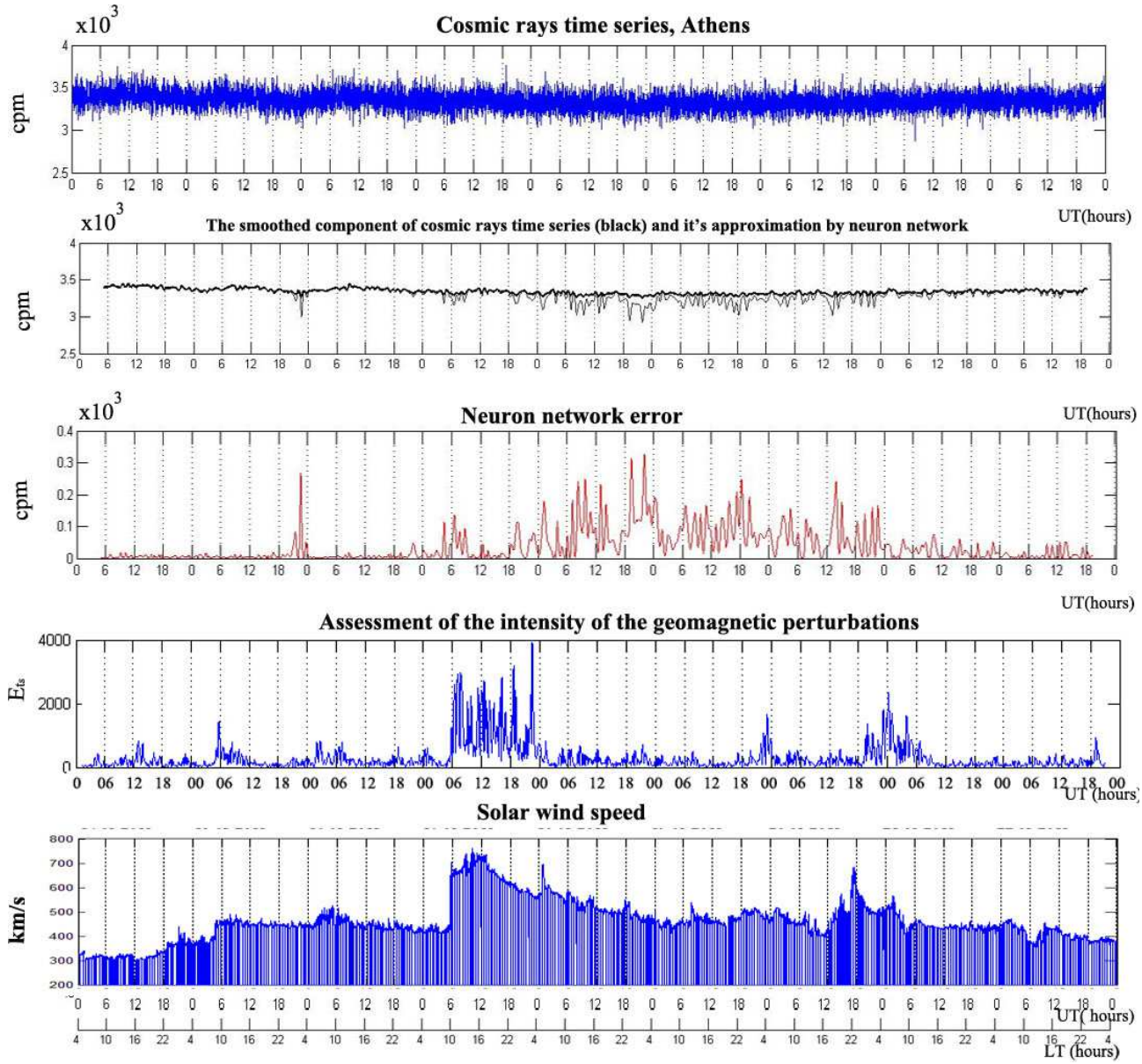


Рис. 9. Результаты анализа данных космических лучей станции «Новосибирск» за период с 14.03.2013 г. по 22.03.2013 г.

[7] *Vecchio A., Laurenza M., Storini M., Carbone V* 2012. New insights on cosmic ray modulation through a joint use of non stationary data-processing methods. *J. Advances Astronomy.* doi:10.1155/2012/834247.

[8] *Chui C. K.* 1992. An introduction in wavelets. New York: Academic Press.

[9] *Daubechies I.* 1992. Ten lectures on wavelets. CBMS-NSF lecture notes ser. Philadelphia: SIAM. 61.

[10] *Mallat S.* 1999. A wavelet tour of signal processing. London: Academic Press.

[11] *Mandrikova O. V., Zalyaev T. L.* 2013. Modeling of the cosmic rays variations on the basis of combination of multiresolution analysis and neural networks with variable structure. *VI Scientific and Technical Conference (International) on Soft Computing and Measurements (SCM'2013) Proceedings.* St. Petersburg. 111–117. (in Russ.)

[12] *Akasofu S. I., Chapman S.* 1972. Solar-terrestrial physics. Oxford University Press, Oxford.

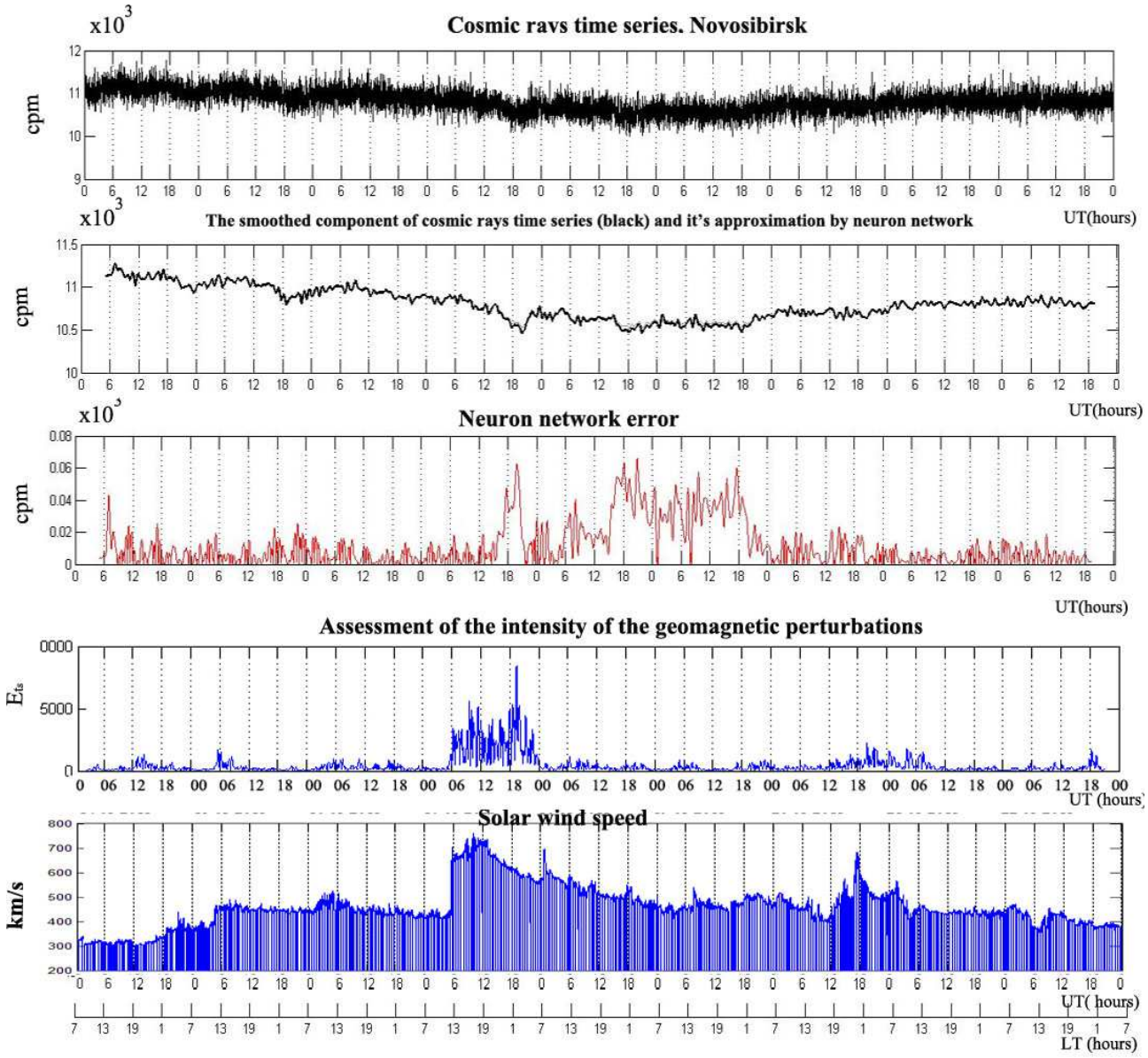


Рис. 10. Результаты анализа данных космических лучей станции «Афины» за за период с 14.03.2013 г. по 22.03.2013 г.

[13] *Haykin S.* 1999. Neural networks: A comprehensive foundation. 2nd ed. New York: Prentice-Hall.

[14] *Kleimenova N. G., Zelinskii N. R., Kozyreva O. V., Malysheva L. M., Solov'ev A. A., and Bogoutdinov Sh. R.* 2013. Pc3 geomagnetic pulsations at near-equatorial latitudes at the initial phase of the magnetic storm of April 5, 2010. *Geomagnetism Aeronomy* 53:313–320.

[15] *Mandrikova O., Solovjev I., Geppenerc V., Taha Al-Kasasbehd R., Klionskiy D.* 2013. Analysis of the Earth's magnetic field variations on the basis of a wavelet-based approach. *Digital Signal Processing.* 23(I.1):329–339.

Моделирование вариативности произношения для уменьшения уровня ошибок при распознавании речи*

***В. Я. Чучупал*[1], *А. А. Коренчиков*[2]**
chuchu@ccas.ru
[1]Москва, Вычислительный Центр им. А. А. Дородницына РАН; [2]Московский государственный университет им. М. В. Ломоносова

Рассматривается возможность снижения уровня ошибок при автоматическом распознавании русской речи за счет использования моделей вариативности произношения. Определена вероятностная модель вариативности произношения, способы оценки ее параметров и реализации в рамках стандартных процедур распознавания речи. Показано, что использование явных моделей вариативности произношения может быть эффективным способом снижения уровня ошибок при распознавании русской разговорной речи, в том числе при несоответствии характеристик обучающего и тестового речевого материала.

Ключевые слова: распознавание речи; акустическое моделирование; вариативность речи; моделирование произношения; скрытые марковские модели

Improving speech recognition accuracy by means of word pronunciation modeling*

***V. J. Chuchupal*[1], *A. A. Korenchikov*[2]**
[1]Dorodnicyn Computing Centre of Russian Academy of Sciences, Moscow; [2]Lomonosov Moscow State University, Moscow

Background: Pronunciation variation modeling evidently has a big potential as a simple way to significantly improve the accuracy of automatic speech recognition. At the same time, the reported improvements in accuracy obtained with pronunciation variation models in experiments are still far from the expected ones.
Methods: The advantages of the so-called explicit pronunciation variation models are explored as an approach for improvement of natural Russian speech recognition accuracy. The probabilistic pronunciation variation model is formally defined as well as the methods of its parameter estimation.
Results: The effect of use of explicit pronunciation variation models is shown to be very dependent on the speech material type. Evaluation on the corpus with Russian read and planned speech shows a negligible effect of using the models. At the same time, the evaluation of pronunciation models on spontaneous Russian speech reveals substantial improvement of automatic speech recognition accuracy.
Conclusions: Despite big promises, there are a lot of efforts necessary to develop pronunciation variation models for speech recognition that will effectively account for speaker and speaking style, accents, and dialects. Nevertheless, right now, the pronunciation model of explicit type can show substantial improvement of recognition accuracy on natural speech recognition task.

Keywords: automatic speech recognition; acoustic modeling; speech variativity; pronunciation modeling; hidden markov models

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-01-00607

Введение

Произнесение слова в системах распознавания слитной речи определяется заданием его произносительной транскрипции: последовательности составляющих это слово фонем. Большинство слов в словаре систем распознавания речи имеет один вариант произношения — каноническую или базовую транскрипцию, которая соответствует нормативному произношению. В повседневной разговорной речи произношение слов может отличаться от нормативного, что является одной из основных причин ошибок при автоматическом распознавании.

Под моделированием вариативности произношения в речевой технологии подразумевают разработку методов определения множества наиболее вероятных акустических образов слов и их последовательностей.

В литературе встречаются два основных подхода к моделированию вариативности произношения [1, 2]. Явное моделирование (explicit modeling) заключается в моделировании вариативности произнесения путем описания возможных изменений в фонемной транскрипции слов [2]. Модель вариативности произношения в данном случае связана с множеством произносительных транскрипций слова. Неявное моделирование (implicit modeling) [3] описывает вариативность произнесения путем изменений в структуре моделей звуков в канонической транскрипции слов, т. е. фонемная транскрипция у слова может быть одна, но имеет сложную форму, например, в виде графа из фонем.

Оба этих подхода не отменяют использования базовых, канонических транскрипций и направлены на определение дополнительных вариантов произнесения слов и словосочетаний.

Использование моделей вариативности произношения имеет высокий потенциал как способ повышения эффективности автоматического распознавания речи. Это очевидно при эвристическом анализе причин появления ошибок и подтверждается данными так называемых модельных экспериментов, когда за счет использования определенных экспертным образом произносительных транскрипций уровень пословной ошибки распознавания — WER (word error rate [4]) может быть уменьшен почти вдвое [5].

Фактические результаты применения моделей вариативности не соответствует ожиданиям. Так в работе [2] на материале корпуса VIOS (для голландского языка) уровень ошибки WER в лучшем случае снизился на 0,8% (с исходных 10,7% до 9,9%), при значительном (4,9 на слово, в среднем) числе допустимых вариантов произнесения. В экспериментах на корпусе данных Switchboard [3] за счет использования неявных моделей вариативности показатель WER снизился на 1,7% (с 39,4% до 37,7%). На корпусе данных NIST 2000 Hub-5 использование моделей вариативности для учета темпа речи дало уменьшение показателя WER на 2,2%: с 54,6% до 52,4% [6].

В приведенных выше работах моделирование, в частности, оценка параметров акустических и языковых моделей, основано на байесовском подходе и использовании моделей порождающего типа. Можно ожидать улучшения результатов при применении дискриминантных методов, которые до недавнего времени в такой задаче не использовались из–за отсутствия достаточных по размеру выборок данных. К настоящему времени для ряда языков уже есть достаточные по объему корпуса данных, собранные, например, в Google и Microsoft.

В работе [7] предложена и исследована дискриминантная модель вариативности произношения при наличии диалекта, в данном случае африканского английского. Вначале экспертным методом были найдены контекстные правила генерации возможных вариантов произнесения по базовым транскрипциям, а затем с использованием т.н. транскрип-

ций от оракула (наиболее корректные из списка N лучших при распознавании), обучалась нейронная сеть для присваивания весов правилам, что позволяло затем вводить веса для получаемых на основе этих правил произносительных вариантов слов.

Поскольку наличие диалекта, как правило, коррелирует с синтаксическими отклонениями, в данной работе одновременно с фонемными транскрипциями проводилась адаптация и языковых моделей.

Теоретически возможное улучшение показателя WER (для транскрипций от оракула) в данном эксперименте было более 10% : с 38,2% до 28,1%. Фактически же внедрение моделей произношения совместно с адаптацией модели языка дало улучшение WER на 2,1%, при этом основной вклад в это улучшение внесла адаптация языковых моделей, модели вариативности произношения принесли 0,6%. Аналогичная модель для оптимизации транскрипций американского стандартного языка также позволила уменьшить WER на 0,8% (с 25,2% до 24,5%).

Наличие огромных выборок данных дает возможность оценить эмпирические частоты транскрипций, что в принципе является оптимальным вариантом при использовании явных методов учета вариативности. В работе [8] на корпусе данных «живого» поиска Windows на мобильных устройствах (Windows Live Search for Mobile voice search task) предложено и экспериментально исследовано несколько способов моделирования вариативности произношения. Рассмотрена эмпирическая модель, когда вероятность фонемных транскрипций слова оценивалась через их относительные частоты, а также нескольких вариантов параметрических моделей со сглаживанием - для возможности оценки вероятности не встречавшихся в обучающей выборке вариантов. Сами варианты произнесения генерировались данным с помощью пофонемного распознавания. В результате удалось (для лучшей из предложенных, линейной комбинации эмпирической и параметрической, моделей) уменьшить величину показателя WER с 34,8% до 33,0%, т. е. на 1,8% или на 5,2%, если рассматривать относительное изменение ошибки.

Приведенные значения показателей эффективности распознавания показывают, что фактические результаты применения моделей вариативности приводят в получению весьма далеких от теоретически ожидаемых результатов, эта ситуация существенно не меняется последние десятилетия [9].

Данная работа во многом вызвана личном опытом авторов. При обработке русской разговорной речи, например выделении ключевых слов, результаты можно заметно улучшить за счет добавления произносительных вариантов для проблемных слов. В представленной работе исследована возможность снижения уровня ошибок автоматического распознавания русской речи за счет использования моделей вариативности произношения. Нас интересовало, насколько можно улучшить результаты распознавания за счет использования явного подхода к моделированию вариативности произношения и в каких случаях. Явный подход был выбран поскольку в этом случае изменения в существующих процедурах распознавания минимальны.

В данном случае мы следовали явному подходу к моделированию вариативности произношения, т. е. выбору наиболее вероятных транскрипций, предполагая, что все изменения в фактическом произнесении можно адекватно описать соответствующими фонемными транскрипциями. Поскольку технологии описания произношения в системах распознавания речи основана на использовании транскрипций, то появление ошибки означает, что фонемные транскрипции корректных слов оказались менее правдоподобными, чем фонемные транскрипции каких-то других слов словаря.

Далее в тексте термины модель произношения слова и его фонемная транскрипция — синонимы, модель вариативности произношения в таком случае соответствует некоторому множеству фонемных транскрипций.

Реализация явного подхода для моделирования вариативности произнесения слов в системе распознавания речи связана с решением следующих задач:

- определение вероятных вариантов произнесения слов словаря,
- оценка параметров модели вариативности,
- результативное использование вариантов произнесения при распознавании.

Модель вариативности произношения

Цель использования модели вариативности произношения в системе распознавания речи — уменьшение числа ошибок распознавания. В данном случае это предполагается достичь за счет нахождения и использования транскрипций, которые больше соответствуют фактически вариантам произнесения, чем базовые.

Различие между словами и транскрипциями заключается в том, что слова относятся к смыслу высказывания, а их произносительные транскрипции определяют акустические параметры и образы слов.

Это различие можно учесть путем детализации формулы классического вероятностного подхода к распознаванию речи [10].

Пусть $X = \{x_t\}, t = 1, \ldots, T$ — наблюдаемый образ в виде последовательности параметров речевого сигнала, а $W = \{w_i\}, i = 1, \ldots, N$ – последовательность слов словаря. Результат распознавания образа X, наиболее вероятную последовательность произнесенных слов W^*, можно определить из уравнения [11]

$$W^* = \arg\max_W P(W|X) = \arg\max_W \frac{P(X|W)P(W)}{P(X)}. \quad (1)$$

Первый сомножитель $P(X|W)$ в числителе (1) соответствует правдоподобию данных при заданной последовательности слов и определяется с помощью акустических моделей. Полученная величина правдоподобия затем умножается на значение $P(W)$, которое определяется с помощью модели языка. Знаменатель $P(X)$ — вероятность наблюдения X, выполняет функции нормализующего члена.

Пусть акустической моделью произнесения некоторого слова w служит его фонемная транскрипция t^w. Множество всех транскрипций слова w обозначим T^w. Моделью произнесения последовательности слов W будет любая последовательность их транскрипций, обозначим все их множество как T^W. Запись t^W будет использоваться для обозначения какой–либо одной последовательности транскрипций из T^W.

Применяемые на практике процедуры распознавания речи и обучения акустических моделей, как правило, определяют лучшую последовательность не самих слов, а их акустических моделей, т. е. вместо (1) фактически используется

$$t^{W^*} = \arg\max_{t^W} \frac{P(X|t^W)P(t^W)}{P(X)}. \quad (2)$$

Наиболее вероятная последовательность слов определяется затем путем отнесения каждой модели соответствующему ей слову, т. е.

$$t^{W^*} \to W^*. \quad (3)$$

Когда слова словаря имеют одну единственную транскрипцию, подходы (1) и (2) очевидно эквивалентны.

Используя равенство $P(t^W) = P(t^W|W)P(W)$, выражение (2) можно записать как

$$W^* = \arg\max_{t^W} \frac{P(X|t^W)P(t^W|W)P(W)}{P(X)}. \tag{4}$$

От выражения (2) выражение (4) отличается наличием члена $P(t^W|W)$, который допускает использование вариативности произношения слов. Множество вероятностей $P(T^W|W) = \{P(t^W|W), t^W \in T^W\}$ можно рассматривать как параметры модели вариативности произношения.

Оценка параметров модели вариативности произношения

Для распознавания речи с использованием критерия (4) нужно знать значения параметров трех моделей: акустической, произносительной и модели языка.

Оценка значений параметров по методу максимальной апостериорной вероятности соответствует использованию критерия

$$P(W|X) = \arg\max_{t^W} \frac{P(X|t^W)P(t^W|W)P(W)}{\sum\limits_{t^W \in T^W} P(X|t^W)P(t^W|W)P(W)} \tag{5}$$

Полученные в результате значения параметров можно рассматривать как дискриминантное решение (5) в том смысле, что оно максимизирует вероятность корректных (для обучающих данных) моделей при минимизации суммарной вероятности всех возможных. Параметры модели языка P(W) в (5), как и для (2), можно считать не зависящими от обучающих акустических сигналов, а их оценку выполнять отдельно на текстовом корпусе данных. Однако параметры моделей произношения $P(T^W|W)$, зависят от акустических обучающих данных, поэтому их независимая от параметров акустических моделей оценка некорректна.

Практическое использование критерия (5) по крайней мере до недавнего времени было довольно затруднительным, поскольку предполагало наличие достаточно больших данных, авторам известны пока единичные подобные эксперименты [7]. Для русского языка подобных корпусов данных, по крайней мере опубликованных, пока нет.

Существенно проще выглядит в данной ситуации оценка параметров модели произношения по методу максимального правдоподобия, т. е. с использованием числителя выражения (5).

Предположим, что, обучающий корпус речевых данных X таков, что для всех высказываний известна не только последовательность слов $w_1 w_2 \cdots w_N$ но и их их моделей $t_1^w t_2^w \cdots t_N^w$. В этом случае наиболее правдоподобная оценка параметров $p(t^w|w)$ определится из выражения:

$$p(t^w|w) = \arg\max_{w,t^w} \prod_{w,t^w} p(t^w|w). \tag{6}$$

Эта оценка аналогична соответствующей оценке для вероятностей появления слов в модели языка [12], т. е. это частота появления соответствующей модели:

$$p(t^w|w) = \frac{\#\{t^w\}}{\#\{w\}}, \tag{7}$$

где символ # означает число событий в фигурных скобках, встретившихся в обучающих данных. Таким образом, наиболее правдоподобной оценкой вероятности появления модели слова является ее относительная частота в обучающей выборке.

Поскольку параметры произносительных и акустических моделей очевидно зависят друг от друга, раздельное независимое оценивание их будет некорректно. Предлагается использовать «покоординатную» оптимизацию: сначала получить максимально правдоподобные оценки по одной группе параметров, полагая другие неизменными, потом то же самое для другой группы параметров и т.д.

Например, полагая первоначально все варианты произнесения слов равновероятными, сначала выполнить с использованием существующих акустических моделей распознавание корпуса данных с ограничением на порядок слов, который известен. Вычислить последовательности наиболее вероятных моделей и оценить частоты всех моделей для каждого слова в соответствии с (7). Затем с использованием определенных последовательностей наиболее вероятных моделей обновить значения параметров акустических моделей. Оба этих этапа чередовать до тех пор, пока перестанут изменяться частота появления транскрипций либо вероятность ошибок распознавания.

Алгоритм вычислений значений параметров представлен ниже.

Алгоритм 1 Алгоритм оценки параметров модели произношения

Вход: Речевой корпус данных
текстовая аннотация корпуса данных,
акустические модели аллофонов,
произносительный (фонемный) словарь
Выход: произносительный фонемный словарь

1: **для всех** высказываний корпуса данных:
2: распознаем высказывание
3: определяем наиболее вероятные транскрипции слов
4: вычисляем границы аллофонов и транскрипций
5: **для всех** слов корпуса данных:
6: оцениваем частоты транскрипций
7: **если** частоты транскрипций изменились **то**
8: переоцениваем параметры акустических моделей аллофонов
9: корректируем акустические модели аллофонов
10: корректируем произносительный словарь
11: на шаг 1
12: **иначе**
13: закончить вычисления

Перед началом работы алгоритма каждое слово обучающей части корпуса данных имеет набор транскрипций, которые соответствуют всем практически возможным вариантам его произношения. Оценка эффективности моделей осуществляется по результатам распознавания на независимой тестовой выборке.

Сходимость алгоритма следует из следующих обстоятельств. На шаге распознавания и переоценки частот транскрипций их количество может только уменьшаться за счет отбрасывания редких вариантов. Шаг переоценки параметров моделей при заданных транскрипциях представляет собой модифицированный применительно к марковским моделям

EM-алгоритм (процедура Баума–Уэлча), т. е. правдоподобие данных гарантированно не уменьшается. Хотя теоретические оценки скорости сходимости EM-алгоритма в подобных задачах авторам не известны, практически процесс сходится достаточно быстро, за 4–6 итераций.

Модификация процедур распознавания речи для учета вариативности произношения

Наиболее известный способ реализации модели вариативности произношения при распознавании речи основан на пополнении произносительного словаря новыми вариантами произнесения слов и распознаванием на основе (2)-(3). Этот подход, однако нельзя рассматривать как наилучшее решение.

Перепишем правую часть равенства (1) в виде

$$P(W|X) = \frac{P(W,X)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X, t^W)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X|t^W)P(t^W)}{P(X)}. \tag{8}$$

Из (4) и (8) следует, что если использовать алгоритм вычислений в соответствии с (2), то определить наиболее вероятную последовательность слов W^* можно из условия

$$W^* = \arg\max_W \sum_{t^W \in T^W} P(t^W|X)P(t^W). \tag{9}$$

Решение в соответствии с (9) определяет наиболее вероятную последовательность слов, а не транскрипций, как (2)-(3), что лучше отвечает интуитивному пониманию решения задачи распознавания: как правило нас интересует, какие слова сказаны, а не то, каким образом они были произнесены.

Алгоритм распознавания с использованием критерия (4) отличается от версии для (2)-(3) тем, что нужно учитывать вероятность появления отдельных транскрипций слова, принимать решение о правдоподобии слова по взвешенной сумме правдоподобий его транскрипций.

Реализация вычислений по (9) потребует дополнительные по сравнению с (2)-(3) шаги для выбора лучшей последовательности слов в соответствии с (9). Поскольку теперь для каждого слова w

$$P(w) = \sum_{t^w \in T^w} P(t^w|X), \tag{10}$$

то (если, например, рассматривать произносительный словарь в виде дерева) в каждом листе дерева нужно вычислить вероятность слова в соответствии с (10).

По сравнению с традиционным (2)–(3) подходом также требуется произвести очевидные изменения в структурах данных процедуры поиска, например отвести память для вероятностей отдельных транскрипций слов и связей между словом и листьями дерева, которые соответствуют моделям этого слова.

Практическая реализация этого алгоритма связана с проблемой, которая возникает из-за процедур обрезки (pruning [12]) вершин дерева лексикона при распознавании. Выражение (10) может включать правдоподобия листьев, которые были выброшены из поиска вследствие их малой вероятности. Необходим альтернативный способ оценки значения выражения (10) для таких случаев.

В связи с этим рассмотрим способ оценки правдоподобия слова, упрощенный вариант (10) с заменой взвешенной суммы правдоподобий моделей на выбор взвешенной максимально правдоподобной модели

$$W^* = \arg\max_{W,t^W} P(t^W|X)P(t^W). \tag{11}$$

В этом случае перечисленные выше практические вычислительные проблемы отсутствуют, а сам алгоритм поиска отличается от общепринятого только наличием «штрафующего» члена $P(t^W|X)$.

Эксперименты и обсуждение результатов

Предложенные модели вариативности произношения сравнивались в ходе численного эксперимента на корпусах данных ISABASE–2 [13] и TeCoRus [14].

Необходимым условием экспериментов с вариативностью произношения является наличие частотности у соответствующих слов в корпусе данных: если слово не встречается или встречается в корпусе данных один раз, сложно судить о вариантах его произнесения. Поскольку цифры и числительные достаточно часто повторяются в речи их, в принципе, естественно было бы использовать в таких измерениях. Недостатком является то, что некоторые цифры короткие и их дополнительных вариантов произнесения может и не быть, а также то, что числительные обычно несут смысловую нагрузку, поэтому произносятся аккуратно, что тоже снижает вариативность.

Обучающая выборка состояла из речевых высказываний 200 дикторов ISABASE–2 (около 40 тыс. предложений) и 50 дикторов TeCoRus (3 тыс. предложений). Составленный по обучающей выборке набор числительных включал 26 слов: цифр от 0 до 9 и числительных до сотни. Этот же набор числительных, включая найденные варианты использовался во всех описанных ниже тестах.

Материал первого теста состоял из данных TeCoRus: 776 слитных цифровых последовательностей, длиной от 2 до 16 цифр, всего 3147 цифр, фактический словарь соответственно был ограничен словоформами цифр. Чтобы оценивать влияние только акустических характеристик на распознавание, модель языка была отключена.

Результаты численного эксперимента, выраженные в терминах пословной ошибки распознавания WER приведены в табл. 1. Здесь и далее в приведенных таблицах колонка «Базовый» содержит результаты для случая использования только базовых фонемных транскрипций, т. е. без вариативности произношения: каждое слово имеет ровно одну, базовую, фонемную транскрипцию . Колонка «Обычный» соответствует методу учета вариативности с использованием (1)–(3), колонка «Опт» – методу учета вариативности на основе (9) и колонка «СубОпт» — методу на основе (11). Значение в строке «Вариат.» характеризует среднюю вариативность произносительного словаря, т. е. среднее число вариантов произносительных транскрипций на слово, которое оценивалась как $\sum_{i=1}^{N} t_i/N$ где N – количество слов в словаре, а t_i – число произносительных транскрипций которые имело i-е слово. В качестве словаря, для которого вычислялась средняя вариативность, рассматривался словарь цифр и числительных.

Результаты приведенные в табл. 1 можно интерпретировать как свидетельство отсутствия вариативности произнесения цифр в данном корпусе. Действительно, дикторы TeCoRus проживали в Москве, имели высшее образование, принадлежали в основном к двум профессиональным группам (в том числе лингвисты), которые аккуратно читали предложенный цифровой материал.

Таблица 1. Показатель пословной ошибки распознавания (WER) для моделей вариативности произношения на данных корпуса TeCoRus

Метод	Базовый	Обычный	Опт	СубОпт
WER	1,62	5,78	2,00	3,17
Вариативный	1,0	1,9	1,9	1,9

Второй эксперимент был проведен на предположительно более вариативном тестовом материале. Обучающий материал был тем же, что и в первом тесте. Тестовый материал включал весь материал первого теста, а также высказывания дикторов TeCoRus, которые содержали числовую информацию (даты, время, номера и т.п.): всего 867 последовательностей цифр или высказываний от 11 дикторов корпуса TeCoRus. Словарь тестовой части включал 129 словоформ, включая 26 цифр и числительных с потенциальной вариативностью. Записи тестовой части, относящиеся к высказываниям с числовой информацией включали потенциально устную речь: дикторы отвечали на вопрос, например, о номерах школ, датах рождения, почтовых индексах и другой персональной числовой информации. Записи в данном случае также включали существенное число различных речевых нарушений. Их было сложно удалить без искажения сигнала. Они служили одним из источников ошибок распознавания.

Таблица 2 содержит результаты измерений уровня пословной ошибки WER в этом случае.

Таблица 2. Показатель WER при использовании различных способов учета вариативности произношения для числовых данных TeCoRus

Метод	Базовый	Обычный	Опт	Субопт
WER	7,78	7,57	7,38	7,44
Вариативный	1,0	1,3	1,3	1,3

Результаты, приведенные в (2), можно рассматривать как более ожидаемые: оптимальным для минимизации показателя WER оказалось использование метода частотного взвешивания произносительных вариантов (9). Метод простого добавления транскрипций (1)–(3) оказался менее эффективным по сравнению как с оптимальным, так и субоптимальным (11), которые учитывают частотность транскрипций, но все же предпочтительнее, чем использование только канонических моделей.

В то же время изменения показателя WER в результате использования моделей произношения представляются незначительными.

Третий эксперимент был проведен на материале естественной разговорной речи. Корпус данных был специально собран и аннотирован для этого эксперимента. Он состоял из фрагментов интервью, которые были взяты с сети Интернет, с сайта радиостанции «Эхо Москвы» [15]. Аудио фрагменты были конвертированы из формата MP3 в формат WAV. Речь была естественная, разговорная в нормальном или быстром темпе. Интервью были предварительно автоматически сегментированы по репликам, найдены и выделены в отдельные файлы фрагменты, которые содержали цифры и числительные.

Полученная таким образом тестовая выборка состояла из 200 коротких реплик с общим словарем в 91 слово, включающим в себя и описанный выше словарь числительных с вариантами произнесения.

Таблица 3 содержит значения показателя WER для этого теста.

Таблица 3. Значения WER для различных моделей вариативности произношения для материала с разговорной русской речью

Метод	Базовый	Обычный	Опт	СубОпт
WER	69,3	57,44	59,7	60,0
Вариативный	1,0	1,3	1,3	1,3

Существенно более высокий уровень абсолютной ошибки распознавания вызван характером речи, использованием кодека, отключением модели языка и несоответствием акустических моделей обучающих и тестовых данных. Кроме того, процедура распознавания и сегментации, которая выделяла словосочетания с числительными из слитной речи не всегда корректно определяла границы слов при слитном произношении, например, в словосочетаниях.

В отличие от предыдущих тестов, где уменьшение относительного уровня ошибки не превышало 5%, в данном случае наблюдается уменьшение относительного уровня ошибок от 13,4% до более чем на 17,1%.

Такие образом для ситуации со словами и словосочетаниями в естественной речи использование моделей вариативности произношения приводит к существенному снижению уровня ошибок распознавания.

Помимо фактически наблюдаемой вариативности произношения возможной причиной снижения уровня ошибок может быть также несоответствие акустических характеристик обучающего и тестового материала, например, из–за использования кодека. За счет этого параметры моделей могли измениться настолько, что это отразилось на наблюдаемых фонемных транскрипциях. Однако это обстоятельство вряд ли играет в данном случае важную роль, поскольку во втором тесте такого эффекта не наблюдалось.

Оптимальный метод формально не оказался лучшим в данном тесте. Этому эффекту можно дать объяснение. Взвешивание решений по транскрипциям уравнивает частоты появления слов. В данном тесте числа имеют в тестовом материале гораздо бóльшую частоту появления, чем другие слова. Использование простого метода пополнения транскрипциями соответствует бóльшей частоте появления для чисел, что соответствует фактическим данным. Таким образом, существенного отклонения от теоретически ожидаемого поведения методов в результатах теста нет.

Заключение

Выполнено исследование методов повышения точности автоматического распознавания русской речи за счет использования явных моделей вариативности произношения. Определена вероятностная модель вариативности произношения, даны способы вычисления ее параметров и способы включения модели вариативности в процедуры распознавания речи. Выполнены численные эксперименты и установлено, что:

— при распознавании т.п. подготовленной читаемой русской речи использование моделей вариативности не является эффективным способом повышения точности распознавания, даже может ухудшать показатели распознавания;

— при распознавании русской спонтанной разговорной речи, в том числе при несоответствии характеристик обучающего и тестового речевого материала, модели вариативности являются эффективным способом снижения уровня ошибок;

— для успешного применения оптимальных моделей вариативности, которые учитывают частоты появления вариантов транскрипций, нужно иметь адекватный тестовому обучающий материал, в противном случае выигрыш от использования таких моделей, по сравнению с традиционными, отсутствует;

Литература

[1] *Fosler-Lussier E.* Dynamic pronunciation models for automatic speech recognition. Ph.D. Thesis. Berkley, CA: University of California, 1999.

[2] *Wester M.* Pronunciation modeling for ASR — knowledge-based and data-derived methods // *Computer Speech Language*, 2003. Vol. 17. P. 69–85.

[3] *Saraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // *Computer Speech Language*, 2004. Vol. 18. No. 4. P. 375–395.

[4] *Word Error Rate* `http://www.echo.msk.ru`.

[5] *Saraclar M., Nock H., Khudanpur S.* Pronunciation modeling by sharing Gaussian densities across phonetic models // *Computer Speech Language*, 2000. Vol. 14. No. 4. P. 137–160.

[6] *Zheng J., Franco H., Stolcke A.* Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // *Speech Communication*, 2003. Vol. 41. P. 273–285.

[7] *Lehr M., Gorman K., Shafran I.* Discriminative pronunciation modeling for dialectal speech recognition // *Proc. Interspeech*, 2014 (in press).

[8] *Hitchinson B., Droppo J.* Learning non-parametric models of pronunciation in automatic speech recognition // *Conference (International) on Acoustics, Speech, and Signal Processing, ICASSP, Proceedings*, 2011. P. 4904–4907.

[9] *Hain T.* Implicit modelling of pronunciation variation in automatic speech recognition // *Speech Communication*, 2005. Vol. 46. P. 171–188.

[10] *Bahl R., Jelinek F., Mercer R. L.* A maximum likilihood approach to continuous speech recognition // *IEEE Trans. Pattern Anal. Machine Intell.*, 1983. Vol. 5. P. 179–190.

[11] *Jelinek F.* Statistical methods for speech recognition. Cambridge, MA: The MIT Press, 1997.

[12] Corpus-based methods in language and speech processing / Ed. by S. Young, G. Bloothooft. Dordrecht: Kluwer Academic Publishers, 1997.

[13] *Богданов Д. С., Кривнова О. Ф., Подрабинович А. Я., Арлазаров В. Л.* Creation of Russian speech databases: Design, processing, development tools // *Conference (International) on Speech and Computers, SPECOM, Proceedings.* Москва, 2004.

[14] *Чучупал В.Я., Маковкин К.А, Чичагов А.В., Кузнецов В.Б., Огарышев В.Ф.* Речевой корпус данных TeCoRus. Свидетельство об официальной регистрации базы данных № 2005620205, 2005.

[15] *Сайт радиостанции Эхо Москвы.* `http://www.echo.msk.ru`.

References

[1] *Fosler-Lussier, E.* 1999. Dynamic pronunciation models for automatic speech recognition. Ph.D. Thesis. Berkley, CA: University of California.

[2] *Wester M.* 2003. Pronunciation modeling for ASR — knowledge-based and data-derived methods. *Computer Speech Language* 17:69–85.

[3] *Saraclar M., Khudanpur S.* 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech Language* 18(4):375–395.

[4] *Word Error Rate* `http://www.echo.msk.ru`.

[5] *Saraclar M., Nock H., Khudanpur S.* 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech Language* 14(4):137–160.

[6] *Zheng J., Franco H., Stolcke A.* 2003. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition. *Speech Communication* 41:273–285.

[7] *Lehr M., Gorman K., Shafran I.* 2014 (in press). Discriminative pronunciation modeling for dialectal speech recognition. *Proc. Interspeech.*

[8] *Hitchinson B., Droppo J.* 2011. Learning non-parametric models of pronunciation in automatic speech recognition. *Conference (International) on Acoustics, Speech, and Signal Processing, ICASSP, Proceedings.* 4904–4907.

[9] *Hain T.* 2005. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Communication* 46:171–188.

[10] *Bahl R., Jelinek F., Mercer R. L.* 1983. A maximum likilihood approach to continuous speech recognition. *IEEE Trans. Pattern Anal. Machine Intell.* 5:179–190.

[11] *Jelinek F.* 1997. Statistical methods for speech recognition. Cambridge, MA: The MIT Press.

[12] *Young S., and G. Bloothooft, eds.* 1997. Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers. 235 p.

[13] *Bogdanov D. S., Krivonva O. F., Podrabinovitch A. J., Arlazarov V. L.* 2004. Creation of Russian speech databases: Design, processing, development tools. *Conference (International) on Speech and Computers, SPECOM, Proceedings.* Moscow. (in Russ.)

[14] *Chuchupal V. J., Makovkin K. A. Chichagov A. V., Kouznetsov V. B., Ogaryshev V. F.* 2005. Speech data corpus TeCoRus. Federal Institute of Industrial Property, RosPatent Database register No. 2005620205. (in Russ.)

[15] *"Echo of Moscow" News/Media WebCite.* `http://www.echo.msk.ru`.

Кластер-анализ пространственных контактов аминокислотных остатков белков с нуклеотидами ДНК*

***Е. Н. Кузнецов*[1], *А. А. Анашкина*[2], *Н. Г. Есипова*[2], *В. Г. Туманян*[2]**

nastya@eimb.ru

[1]Институт проблем управления им. В. А. Трапезникова РАН, Россия, Москва 117997, ул. Профсоюзная, 65; [2]Институт молекулярной биологии им. В. А. Энгельгардта РАН, Россия, Москва 119991, ул. Вавилова, 32

Предлагается классификация аминокислотных остатков по признакам контактов аминокислот белков с нуклеотидами ДНК. Аминокислотные остатки обладают множеством различных свойств и функций и могут одновременно принадлежать к разным классам, поэтому в работе рассматриваются классификации с разными типами размытости. Для определения количества и площади контактов каждой аминокислоты с каждым нуклеотидом в 1937 комплексах использовали разбиение Вороного–Делоне. Задача классификации аминокислотных остатков с разными типами размытости решалась с помощью общего вариационного подхода. Было показано, что около 30% всех контактов между аминокислотами и нуклеотидами в комплексах белок–ДНК являются неслучайными. Методами четкой классификации показано существование инвариантов кластеризации аминокислот. Методами размытой классификации показано, что классификация аминокислот на шесть классов является оптимальной для задачи белок-нуклеинового распознавания.

Ключевые слова: кластер-анализ; размытая классификация; контакты аминокислота-нуклеотид; разбиение Вороного–Делоне; свойства аминокислотных остатков

Cluster analysis for spatial contacts of amino acid residues of proteins with DNA nucleotides*

***E. N. Kuznetsov*[1], *A. A. Anashkina*[2], *N. G. Esipova*[2], *and V. G. Tumanyan*[2]**

[1]Trapeznikov Institute of Control Sciences RAS, 65 Profsoyuznaya Str., Moscow 117997, Russia; [2]Engelhardt Institute of Molecular Biology RAS, 32 Vavilov Str., Moscow 119991, Russia

Background: Amino acids are classified on the basis of protein–DNA contacts geometry and statistics. Amino acid residues have a variety of properties and can simultaneously belong to different classes. So, it was interesting to use the classification of amino acids with different types of fuzzing.

Methods: Voronoi–Delaunay tessellation was used to determine the spatial relationship between the amino acids of proteins and DNA nucleotides from 1937 protein–DNA complexes. General variation approach was used for the classification of amino acids with different types of fusion.

Results: It was shown that about 30% of all contacts between amino acids and nucleotides in protein–DNA complexes are not random. Crisp classification methods showed the existence of clustering invariants of amino acids at the lowest level of association. It was shown by fuzzy classification methods that six classes are optimal for protein–DNA recognition task.

Concluding Remarks: Fuzzy classification of amino acids data can be used to construct the substitution matrix for DNA-binding protein sequences and protein-DNA binding analysis.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 14-04-00639-а и 12-07-00634-а.

Keywords: *cluster analysis; crisp classification; fuzzy classification; protein–DNA interactions*

Введение

Проблема специфичности взаимодействия ДНК-белок лежит в основе понимания механизмов экспрессии генов, а, следовательно, механизмов реализации генетической информации на различных уровнях строения биообъектов. Различают специфическое и неспецифическое связывание нуклеиновых кислот белком: под первым понимается избирательное взаимодействие определенного участка нуклеиновой кислоты с определенным белком, под вторым — равновероятное взаимодействие белка с различными последовательностями нуклеиновых кислот в различных участках генома [1, 2].

Из анализа первых рентгеновских структур белок-нуклеиновых комплексов стало очевидно, что в создание комплекса вносят свой вклад множество различных факторов: водородные связи, опосредованные водой контакты, взаимные конформационные перестройки, изгибы и искажения, высвобождение ионов, электростатика, Ван дер Ваальсовые взаимодействия, гидрофобный эффект [3, 4, 5].

Вычислительные методы, опирающиеся на кристаллографические исследования, широко и успешно использующиеся для оценки энергии взаимодействий белок-лиганд [6, 7, 8, 9], должны быть применимы и для понимания формирования белок-ДНК комплексов. Исследователи [10, 11] пытались оценить вклад каждой пары аминокислотный остаток/нуклеотид в общую аффинность белка к ДНК. Другой подход, предложенный [12], предполагал, что общая оценка, отражающая комплементарность между белком и его специфической ДНК, может быть вычислена методами статистического анализа частоты взаимодействия между парой аминокислотный остаток/нуклеотид, таким образом подразумевая аддитивность в энергии связывания. Другие попытки качественно или количественно описать взаимодействие между белком и ДНК [13, 14, 15, 16, 17, 18, 19, 20, 21] также опираются на доступные трехмерные кристаллические структуры белков, связанных с ДНК. Таким образом, все многообразие информации о правилах, управляющих биомолекулярным распознаванием, получено из структурных данных, в основном из рентгеноструктурного анализа и ЯМР.

Белок и ДНК различаются структурно и химически. В комплексах белок-ДНК молекулярные интерфейсы пространственно комплементарны, и распознавание является точным структурным процессом. Стереохимическая ориентация взаимодействующих поверхностей партнеров определяет комплементарность химических контактов и неизбежно влечет за собой существование молекул с комплементарными водородными донорными и акцепторными группами. Это означает химическое распознавание.

В данной работе мы задались целью найти способ классификации аминокислот, наиболее интегрально учитывающий факторы, определяющие образование специфических комплексов ДНК-белок. Известны различные классификации аминокислотных остатков, основанные, в частности, на их физико-химических свойствах [22, 23], на анализе точечных мутаций и кластеризации матриц замен [24], на анализе соседних по последовательности аминокислотных остатков [25] и т. д. При этом используется большое разнообразие методов кластер-анализа и автоматической классификации, в том числе методы иерархической классификации [26, 27], методы типа k-средних, вариационные методы классификации, методы многомерного шкалирования [22] и др. Очевидно, что универсальной классификации

аминокислот не существует, и каждая классификация предназначается для целей определенного исследования [28]. Это означает, что имеет смысл говорить о контекст-зависимой классификации для решения конкретной задачи.

Для поиска конкретных способов реализации белок-нуклеинового узнавания авторы решили создать классификацию аминокислот на основе анализа геометрических характеристик структур комплексов белок-ДНК. Аминокислотные остатки в составе белков, взаимодействующих с ДНК, образуют пространственные контакты с нуклеиновыми основаниями и сахарофосфатным остовом ДНК. Был проведен анализ пространственного взаимного расположения аминокислотных остатков и нуклеотидов на большой выборке комплексов белок-ДНК (1937 комплексов, т. е. все известные структуры белок-ДНК в базе данных Protein Data Bank на момент исследования). Для расчета количества и площади контактов использовался подход, основанный на пространственном разбиении Вороного–Делоне [29, 30].

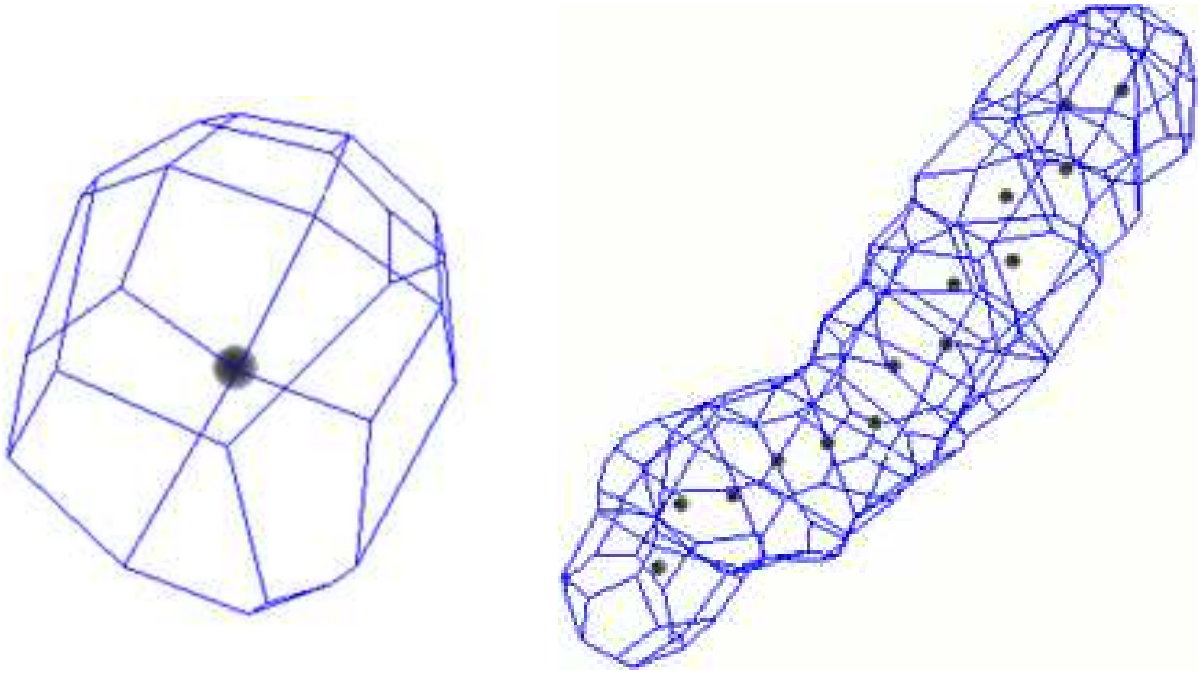
Для проверки надежности предлагаемого подхода к решению общей проблемы белок-нуклеинового узнавания доступная выборка пространственных структур белок-ДНК была разбита на две подвыборки в 987 и 950 комплексов. Было показано совпадение результатов классификаций для каждой подвыборки и выборки в целом для иерархических методов классификации, а для вариационных – совпадение с точностью до задания начальных условий.

В данной работе впервые (1) в основу классификации аминокислот положены геометрические характеристики структур комплексов белок-ДНК; (2) для конкретного представления пространственного взаимодействия аминокислотных остатков белков и нуклеотидов ДНК использовано разбиение Вороного-Делоне; (3) в качестве признаков для применения методов кластер-анализа использованы как статистика контактов, так и статистика площадей контактов между аминокислотными остатками и нуклеотидами белок-нуклеиновых комплексов.

Разбиение Вороного-Делоне

Для любого центра из системы центров можно указать область пространства, все точки которой ближе к данному центру, чем к любому другому центру системы. Такая область называется многогранником Вороного или областью Вороного. Разбиение Вороного разделяет пространство между набором центров. Каждый центр системы посредством граней многогранника Вороного определяет своих геометрических соседей. Те, в свою очередь, определяют своих соседей и т. д. Таким образом, можно говорить о графе, вершинами которого являются центры системы, а связность определена через геометрическое соседство. В трехмерном пространстве область Вороного для произвольного центра системы является выпуклым многогранником (полиэдром). Области Вороного для каждого центра системы образуют «сеть» полиэдров, называемую разбиением Вороного [29]. Если внутри описанной сферы тетраэдра, определенного четырьмя центрами, нет других центров системы, такой тетраэдр называется симплексом Делоне. Совокупность всех симплексов Делоне системы заполняет пространство без наложений и щелей, т. е. подобно многогранникам Вороного реализует разбиение пространства, но на этот раз на тетраэдры. Это разбиение называется разбиением Делоне. Как методом пустого шара Делоне, так и с помощью плоскостей Вороного мы выявляем одну и ту же систему центров. Итак, можно говорить о едином разбиении Вороного–Делоне, в котором мы видим одновременно как мозаику многогранников Вороного, так и симплексов Делоне. Каждый симплекс Делоне соответствует определенной вершине Вороного, и, наоборот, каждой вершине Вороного со-

ответствует симплекс Делоне. Эти разбиения являются дуальными, и являются топологически эквивалентными. Таким образом, данный метод распределяет пространство внутри белковой глобулы между всеми ее атомами по следующему принципу: разделяющая плоскость проводится между двумя соседними атомами через середину отрезка, соединяющего эти атомы и перпендикулярно ему. Такие плоскости образуют вокруг каждого атома выпуклый многогранник произвольного вида, называемый полиэдром Вороного (рис. 1). Область внутри многогранника лежит ближе к данному атому, чем к любому другому. Таким образом, контакт между двумя атомами существует, если у этих атомов есть общая грань полиэдра Вороного с площадью, отличной от нуля. Следовательно, контакт между двумя аминокислотами определяется как совокупность общих граней полиэдров Вороного составляющих их атомов. Площадь такого контакта определяется как сумма площадей граней составляющих его атомарных контактов.



(а) А. Полиэдр Вороного для одного атома (б) Полиэдры Вороного для цепочки атомов

Рис. 1. (*а*) Полиэдр Вороного, построенный вокруг одного атома. В общем случае он является выпуклым многогранником с произвольным числом граней разного размера, зависящем от расположения соседних атомов. Атомы-соседи не показаны. (*б*) Разбиение Вороного цепочки атомов. Атомы-соседи, окружающие цепочку, не показаны

С помощью программы, реализующей трехмерное разбиение Вороного-Делоне для координат атомов структур в формате PDB, исследовали полученные на основе данных рентгеноструктурного анализа комплексы ДНК-белок. При отборе рассматривали только структуры, содержащие одновременно как белковые цепи, так и ДНК, и исключали структуры, содержащие РНК или ДНК/РНК-гибриды. Всего исследовали 1937 структур. Контакты между белками и ДНК были вычислены на основе анализа координат атомов пространственных структур белок-ДНК методом разбиения Вороного–Делоне [31]. Помимо информации о контактах, в результате применения этого метода мы имеем данные о площади общей грани полиэдров соседних атомов. Таким образом, результатом прове-

денного разбиения Вороного–Делоне являются таблицы контактов как между атомами аминокислот и атомами нуклеотидов, так и между более крупными пространственными единицами — аминокислотными остатками и нуклеотидами, как по числу контактов, так и по суммарной площади. Ранее мы применили это разбиение для анализа белок-белковых и белок-нуклеиновых взаимодействий [29, 30]. Программа для построения разбиения написана на языке C++, ее исходный код доступен по запросу авторам статьи через электронную почту.

Модели случайно и неслучайно контактирующих химических единиц (аминокислот/нуклеотидов)

Для полноценной интерпретации полученных данных нам необходимо опираться на статистическую математическую модель контактирующих аминокислот/нуклеотидов, для того, чтобы оценить и выявить отклонения от случайных явлений. Сделаем несколько принципиальных приближений. Первое состоит в том, что мы будем рассматривать область взаимодействия белковых (белок-нуклеиновых) молекул как поверхность, образованную гранями полиэдров Вороного пар атомов, один из которых принадлежит одной молекуле, а другой – второй молекуле. Второе предположение заключается в том, что все контакты на уровне аминокислотных остатков (остаток/нуклеотид) можно рассматривать как совокупность случайных и неслучайных контактов. Под неслучайными, специфическими контактами подразумеваются контакты, возникающие на участках пространственных структур между определенными химическими группами и/или вследствие определенных типов взаимодействий, сопровождающихся выигрышем в энергии, между элементами на определенных местах в структурах. Такие контакты могут иметь характерную площадь контакта (или несколько, в случае нескольких возможных взаимных расположений). Случайные контакты, в свою очередь, образуются как следствие пространственного сближения двух остатков по причине формирования неслучайных контактов. Таким образом, каждый тип контакта между двумя аминокислотными остатками, например Arg-Glu, может образовывать как случайные, так и неслучайные контакты. В этом приближении оценим распределение площади контакта между остатками на поверхности белок-белкового интерфейса.

Случайные контакты

Предположим, что два круга бросают на некоторую область случайным образом, и каждый раз фиксируют площадь перекрывания. Для упрощения предположим, что эти круги одинаковые с радиусом r, а бросание производится на квадратную область с длиной стороны, равной R. Площадь каждого круга πr^2, тогда площадь их перекрывания лежит в диапазоне $[0, \pi r^2]$. Определим зависимость площади пересечения S от расстояния между центрами кругов L. Очевидно, что если $L \geqslant 2r$, то $S = 0$. Требуется вычислить площадь сектора $AOBs$ и площадь треугольника AOB для вычисления площади сегмента, ограниченного дугой s и отрезком AB. Площадь пересечения кругов будет:

$$S(L) = 2\left(r^2 \arcsin\left(\sqrt{1 - \frac{L^2}{4r^2}}\right) - \frac{L\sqrt{r^2 - L^2/4}}{2}\right) \tag{1}$$

где L расстояние между центрами кругов. Эта формула верна, если $L \in [0, 2r]$.

Пусть координаты центров кругов (x_1, y_1) и (x_2, y_2) соответственно. Тогда расстояние между центрами кругов $l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Если мы поместим область для

бросания в начале координат, то координаты центров кругов могут принимать значения из интервала $[r, R-r]$. Найдем вероятность того, что площадь пресечения S упавших кругов равна нулю. Выражаясь математическим языком, расстояние между центрами кругов должно быть больше или равно суммы их радиусов:

$$l \geqslant 2r\,. \tag{2}$$

Как известно, такая вероятность равна отношению объема пространства, удовлетворяющего этому условию, ко всему объему пространства, которое могут принимать значения координат центров кругов. Объем пространства, которое могут принимать значения координат центров кругов $(R-2r)^4$. Для вычисления объема пространства, в котором площадь пересечения кругов будет равна нулю, произведем замену $u_1 = (x_1 - x_2)/\sqrt{2}$, $u_2 = (x_1 + x_2)/\sqrt{2}$, $v_1 = (y_1 - y_2)/\sqrt{2}$, $v_2 = (y_1 + y_2)/\sqrt{2}$.

Тогда матрица перехода будет выглядеть так:

$$\begin{pmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix}\,.$$

Неравенство (2) можно теперь записать в виде

$$u_1^2 + v_1^2 \geqslant 2r^2\,, \tag{3}$$

где u_1 и v_1 могут принимать значения из интервала $[(-R+2r)/\sqrt{2}; (R-2r)/\sqrt{2}]$. Заметим, что неравенство (3) представляет собой пространство вне круга радиусом $\sqrt{2}r$ и внутри квадрата со стороной $R-2r$. Таким образом, вероятность того, что площадь пересечения кругов будет равна нулю

$$P = \frac{((R-2r)^2 - 2\pi r^2)(R-2r)^2}{(R-2r)^4} = \frac{((R-2r)^2 - 2\pi r^2)}{(R-2r)^2} = 1 - \frac{2\pi r^2}{(R-2r)^2}\,. \tag{4}$$

Таким образом, вероятность того, что расстояние между центрами кругов больше L, выражается формулой:

$$P = 1 - \frac{\pi L^2}{2(R-2r)^2}\,. \tag{5}$$

Для нахождения плотности вероятности dP/dS от площади пересечения S можно переписать уравнение в параметрическом виде, поскольку нельзя выразить L как функцию от S в явном виде:

$$\left.\begin{aligned} \frac{dP}{dS} = \frac{dP}{dL}\frac{dL}{dS} = \left(-\frac{\pi L}{(R-2r)^2}\right)\left(-\frac{1}{2\sqrt{r^2 - L^2/4}}\right) = \frac{\pi L}{r(R-2r)^2\sqrt{1 - L^2/(4r^2)}}\,; \\ S(L) = 2\left(r^2 \mathrm{acrsin}\left(\sqrt{1 - \frac{L^2}{4r^2}}\right) - \frac{Lr}{2}\sqrt{1 - \frac{L^2}{4r^2}}\right)\,. \end{aligned}\right\} \tag{6}$$

Зависимость (6) показана на рис. 2, кривая А. По мере увеличения площади контакта число контактов резко уменьшается. Следовательно, среднее распределения близко к нулю. Другими словами, из распределения для случайных контактов видно наличие большого числа малых по площади контактов.

Неслучайные контакты

Логично предположить, что специфические контакты обладают некоторой, отличной от нуля средней площадью контакта, обусловленной физико-химической природой взаимодействия остатков. Предположим, что специфические взаимодействия стремятся образовать максимально большой возможный контакт. В этом случае распределение расстояний между центрами кругов подчиняется нормальному распределению, напоминая задачу о стрельбе по мишени. Выразим распределение площадей неслучайных контактов также в параметрическом виде:

$$\left.\begin{aligned} f(L) &= \frac{1}{\sigma\sqrt{2\pi}} e^{(L-a)^2/(2\sigma^2)}; \\ S(L) &= 2\left(r^2 \operatorname{acrsin}\left(\sqrt{1-\frac{L^2}{4r^2}}\right) - \frac{Lr}{2}\sqrt{1-\frac{L^2}{4r^2}}\right). \end{aligned}\right\} \tag{7}$$

Зависимость (7) показана на рис. 2, кривая B. Кривая имеет куполообразную форму, несимметричная, с некоторым, существенно отличным от нуля средним значением. Распределение для специфических контактов отражает существование некоторой характерной площади контакта. В общем случае уравнения (6) и (7) должны входить в суммарное уравнение, отражающее общее распределение, с некоторыми весовыми функциями, отражающими пропорцию между специфическими и случайными контактами. Площадь под суммарной кривой должна равняться 1.

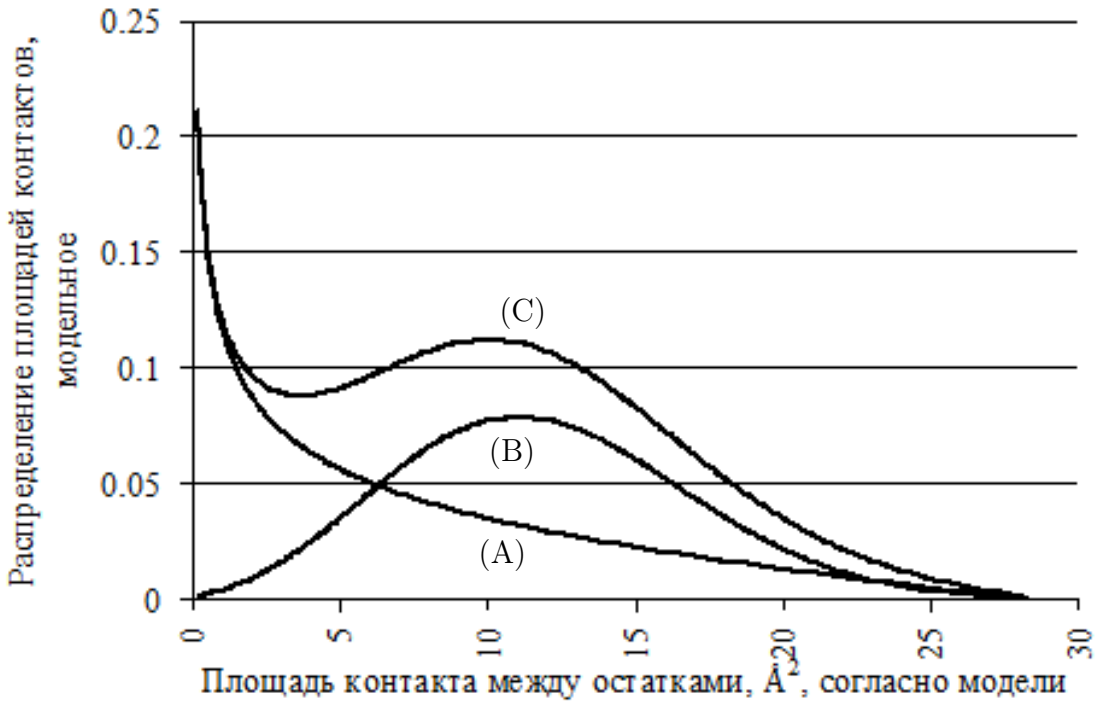


Рис. 2. Графики, отражающие системы (6) и (7), моделируют распределения площадей случайных (A) и специфических (B) контактов: A —график системы (6) в параметрической форме. График отражает распределение площади случайных контактов; B — график системы (7) в параметрической форме. График отражает распределение площади специфических контактов; C — сумма графиков A и B. Параметры, использованные в данном случае: $R = 20$, $r = 3$, $a = 3$, $\sigma = 1$

Классификация аминокислотных остатков на основе сравнительного анализа контактов в структурах комплексов белок-ДНК и специфические взаимодействия ДНК-белок

Белок-нуклеиновое распознавание представляется сложным многоступенчатым процессом, и найти соответствие между типами аминокислотных остатков и типами распознаваемых ими нуклеиновых оснований, т. е. так называемый «код» ДНК-белкового узнавания, было и остается мечтой множества исследователей. Спустя годы поисков стало понятно, что простого, единственного кода белок-нуклеинового узнавания не существует. Существует ли вырожденный код или несколько таких кодов, когда одной группе нуклеотидов соответствует определенная группа аминокислотных остатков? Чтобы развить подход к решению такого сложного вопроса, мы задались целью найти способ классификации аминокислот, наиболее интегрально включающей признаки, определяющие образование специфических комплексов ДНК-белок. Известны различные классификации аминокислотных остатков, основанные, в частности, на их физико-химических свойствах, на анализе точечных мутаций, на анализе соседних по последовательности аминокислотных остатков, кластеризации матриц замен и так далее. Это означает, что имеет смысл говорить о контекст-зависимой классификации для решения конкретной задачи. В нашем случае, для поиска способов реализации белок-нуклеинового узнавания, мы решили создать классификацию аминокислот на основе анализа геометрических характеристик структур комплексов белок-ДНК. Аминокислотные остатки в составе белков, взаимодействующих с ДНК, образуют пространственные контакты с нуклеиновыми основаниями и сахарофосфатным остовом ДНК. Для построения независимой классификации аминокислотных остатков, наилучшим образом применимой для установления вырожденного кода узнавания белком ДНК, можно использовать статистику контактов аминокислот с нуклеотидами. Мы провели анализ поведения аминокислотных остатков по отношению к нуклеотидам на основе представительной статистики, полученной нами в данной работе с помощью разбиения Вороного–Делоне. При этом впервые в основу классификации аминокислот положены как статистика контактов, так и статистика площадей контактов между аминокислотными остатками и нуклеотидами белок-нуклеиновых комплексов. Статистика контактов и площадей контактов аминокислота/нуклеотид, полученная в данной работе на выборке из 1937 белок-ДНК комплексов методом Вороного–Делоне, представлена в табл. 1.

Эта статистика является промежуточным результатом в рамках способа классификации аминокислот применительно к процессам белок-нуклеинового взаимодействия. Числа в таблице отражают количество случаев (число событий), когда аминокислота, соответствующая строке, образует контакт (пространственно сближена) с нуклеотидом, соответствующим столбцу. Определение сходства аминокислотных остатков путем анализа матриц контактов и площадей контактов. Измерение близости между аминокислотами мы проводили на основе сравнения соответствующих строк в матрице контактов. В качестве меры близости (сходства) строк были использованы девять мер расстояния: D_1–D_4, D_7–D_{10}, D_{12} [32]. Далее анализ расстояний между соответствующими строками матрицы контактов проводили при помощи различных методов кластеризации. Иерархические методы кластеризации включали метод ближней связи, дальней связи и метод средней связи. Неиерархические методы включали метод k-средних и метод Уорда с заданным числом классов от четырех до семи и с разными критериями кластеризации: Trace(W), Trace(W)/Median и Wilks' Lambda, где W — внутрикластерная ковариационная матрица [33].

Таблица 1. Количество и суммарные площади контактов между аминокислотными остатками белка и нуклеотидами ДНК в белок-нуклеиновых комплексах. Данные получены с помощью анализа пространственных координат атомов аминокислотных остатков и нуклеотидов в 1937 комплексах белок-ДНК методом разбиения Вороного–Делоне

	Количество контактов, шт.				Площади контактов, Å^2			
	A	T	G	C	A	T	G	C
ALA	2408	2764	2553	2461	16966,05	24384,55	18979,93	20319,00
ARG	11039	11319	12667	9013	134455,83	138697,44	166185,09	100840,31
ASN	3285	3936	3275	2980	33582,22	40044,17	32957,42	28341,64
ASP	1376	1065	2060	1747	10820,57	7528,08	15287,42	13140,64
CYS	328	352	340	341	2750,04	3128,09	1968,50	3369,63
GLN	2959	2802	2687	2720	29097,04	27407,50	30136,37	28434,73
GLU	1702	1776	2037	2079	12592,60	13869,09	16019,02	16700,26
GLY	3561	4144	3597	3494	27282,42	35127,11	29074,57	27561,70
HIS	1895	2372	1951	1356	19315,16	24701,81	21231,09	12459,42
ILE	2004	2169	2026	1790	17583,49	21376,97	21771,69	14961,05
LEU	1907	2203	1812	1698	15658,60	22519,82	17680,12	14674,31
LYS	7964	8156	8184	7123	79052,97	83339,78	85176,25	67237,77
MET	741	1128	1008	742	7884,58	12799,33	10824,30	8368,55
PHE	1458	1847	1643	1461	20434,49	25979,69	19290,77	17668,08
PRO	1905	2070	1610	1586	17408,12	17352,59	12599,16	11623,36
SER	3998	4897	4596	3496	37826,26	52214,91	44773,53	31770,22
THR	4066	4902	4095	3517	40021,98	54137,06	40657,42	34666,27
TRP	544	625	680	790	7120,88	10196,77	8758,85	10829,64
TYR	2476	2906	2992	2389	29171,89	39001,30	38277,34	31241,26
VAL	2263	2483	2097	1733	23049,29	20393,23	18528,68	14429,04

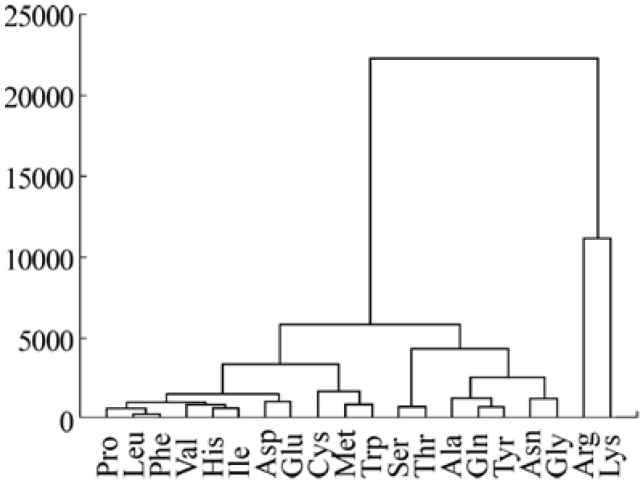


Рис. 3. Иерархическое дерево, полученное в результате применения метода средней связи и расстояния D_1 к данным табл. 1 о количестве контактов аминокислотных остатков белков с нуклеотидами ДНК, рассчитанным при помощи разбиения Вороного–Делоне

В результате применения девяти методов оценки расстояний и трех иерархических методов кластеризации было получено 27 иерархических деревьев, отражающих структуру взаимосвязи между аминокислотными остатками на основе пространственного взаимодействия аминокислотных остатков с нуклеотидами ДНК. На рис. 3 для примера приведено иерархическое дерево, полученное в результате применения кластеризации по методу средней связи и расстояния $D1$ (манхэттенское расстояние или «сити-блок»). Во всех 27 случаях было найдено, что следующие аминокислоты группируются в пары: Leu и Phe; Ser и Thr; Met и Trp; Asn и Gly; Gln и Tyr. His и Ile группируются в пары в 19 случаях из 27, Arg и Lys в 21 случае из 27, в 20 случаях из 27 Cys входит в группу с Met и Trp. Ala входит в группу Gln и Tyr в 25 случаях. В 17 случаях из 27 образуется группа из аминокислот Leu, Phe, Pro, His, Ile, Val. В остальных случаях все эти шесть аминокислот располагаются в одном классе, дополняясь глютаминовой кислотой.

По результатам иерархической кластеризации можно выделить шесть классов аминокислот: I. Leu, Phe, Pro, His, Ile, Val. II. Asp, Glu. III. Met, Trp, Cys. IV. Ser, Thr. V. Gln, Tyr, Asn, Gly, Ala. VI. Arg, Lys. Главным физико-химическим свойством, объединяющим аминокислоты класса I, является большое количество неполярных групп, входящих в боковые радикалы этих аминокислот. Это характерно и для гистидина, невзирая на его заряд. Класс II наблюдается только в девяти случаях из 27, в остальных случаях отрицательно заряженные аспарагиновая и глютаминовая кислоты могут присоединяться к классу I или оставаться в фоновом классе. В классе III аминокислоты являются неполярными. В девяти случаях из 27 цистеин образует свой собственный класс. Обратим внимание на то, что метионин и цистеин являются серосодержащими аминокислотами, а триптофан содержит ароматическое кольцо. Таким образом, аминокислоты этого класса обладают большими боковыми радикалами и поэтому занимают значительное пространство в интерфейсе белок-ДНК. Класс IV образуют аминокислоты, обладающие очень близкими физико-химическими свойствами: содержат одинаковые функциональные группы и имеют одинаковую длину бокового радикала. Класс V интересен в нескольких аспектах. Во-первых, он содержит всегда составляющие пару две разных аминокислоты — аспарагин и глицин. Причем аспарагин и глицин образуют пару при использовании любого способа вычисления расстояния и применении любого иерархического метода кластеризации. Именно эти аминокислоты могут принимать конформации, запрещенные для остальных аминокислот. Например, они входят в состав некоторых бета-изгибов II-типа [34]. Аминокислоты этой группы (глутамин, аспарагин, аланин) часто входят в состав левой спирали типа PPII. Класс VI включает в себя аргинин и лизин, которые в некоторых случаях образуют отдельные классы, и поэтому могут являться самостоятельными объектами для возникновения кодовых комбинаций, важных при развитии ДНК-белкового узнавания. Мы приводим ниже только несколько примеров результатов неиерархической классификации, допускающих интерпретацию и определенное сравнение с иерархическими методами. Например, в результате анализа евклидового расстояния (D_2) методом средней связи Кинга и методом Уорда мы получили, что образуется четыре класса: A. Ala, Gln, Asn, Ser. B. Arg. C. Asp, Glu, His, Cys. D. Lys. Остальные аминокислоты остаются в фоновом классе. Однако это искупается возможностью ясной интерпретации выявленных классов. Класс A содержит аминокислоты, соответствующие левой спирали типа PPII, о которой мы говорили при обсуждении класса V иерархической кластеризации. Класс B характерен для протаминов, класс C – для альфа-спиральных структур, а класс D — для гистонов. Методом к-средних Мак-Куина частично воспроизведены результаты иерархических методов: 1. Gln, Tyr, Asn, Gly, Ser, Thr. 2. Arg, Lys. 3. Ala, Asp, Glu, Met, Trp, Cys, Leu, Phe,

Pro, His, Ile, Val. Здесь первый класс, по сути, объединяет классы четыре и пять, полученные иерархическими методами классификации, второй класс есть класс шесть, а третий класс объединяет классы 1–3. Исключение составляет лишь аланин. Метод k-средних, если в качестве критерия кластеризации взять нормированную суммарную внутриклассовую дисперсию (Trace(W)/Median), с заданным числом классов от трех до шести, дает противоречивые результаты. Очевидно, что существует не один, а несколько близких по эффективности способов группирования аминокислот в контексте определенной проблемы, при этом признаки, определяющие классификацию, могут быть непосредственно не связанными с их физико-химическими свойствами. Поэтому кластеризация, созданная на основе признаков, выявленных при ДНК-белковом узнавании, не будет адекватной, если мы попытаемся использовать ее в рамках проблемы узнавания белок-белок. Используя различные методы оценки расстояния и способы объединения аминокислот в группы, можно выявить инварианты кластеризации аминокислот. Результаты, полученные с помощью иерархических методов кластеризации, имеют общие характерные черты. Надо еще раз подчеркнуть, что следующие аминокислоты группируются в пары вне зависимости от способа вычисления расстояния и метода кластеризации: Leu и Phe; Ser и Thr; Met и Trp; Asn и Gly; Gln и Tyr. Из пяти пар бинарной классификации только две пары находят четкое физико-химическое и структурное толкование, но все пять пар связаны с различными типами локальных структур полипептидной цепи. Дополнительно отметим подобие структур лейцина и фенилаланина. На верхних уровнях организации состав классов также практически неизменен. Физические свойства аминокислот, такие как гидрофобность, заряд, наличие гидроксильной группы, проявляют себя во взаимодействиях с ДНК не в полной мере, что отражается на классификации этих аминокислот. Сходства химической структуры боковых радикалов также оказалось недостаточно для разделения аминокислот по группам. Любопытно выглядит объединение в один класс таких разных по физико-химическим свойствам аминокислот, как глютамин, аспарагин, тирозин, глицин и аланин. Метионин, триптофан и цистеин, образующие класс III, также обладают очень разными физико-химическими свойствами. Метионин и цистеин являются серосодержащими аминокислотами, в то время как триптофан имеет большую ароматическую группу. Цистеин в семи классификациях из 27 образует собственный класс. В физико-химическом смысле он не имеет аналогов среди аминокислот.

Вариационный подход к задаче классификации аминокислотных остатков

Проведенный выше классификационный анализ аминокислот, с нашей точки зрения, не полностью описывает все многообразие их свойств. Кроме того, описанные выше методы не позволяют изучить группировку аминокислот в матрицах эволюционных замен. Эти матрицы характерны тем, что замена аминокислоты на аминокислоту того же типа характеризуется некоторым, отличным от нуля, числом. Таким образом, нарушается требование, что сходство объекта с самим собой абсолютно. Для решения этой задачи мы воспользовались общим вариационным подходом к задаче классификационного анализа. Общий вариационный подход к задаче классификационного анализа формулируется при помощи четырех основных категорий: классифицируемое множество объектов, класс допустимых классификаций, способ описания класса и функционал качества разбиения [35].

1. Классифицируемое множество объектов. В нашей задаче классифицируемое множество объектов состоит из $N = 20$ типов аминокислотных остатков. Обозначим это множе-

ство как $X = \{x_1, ..., x_n\}$. Каждый объект i описывается через коэффициенты матрицы замен аминокислот.

2. Класс допустимых классификаций. Пусть требуется разбить множество объектов на K классов. Обозначим принадлежность любого объекта i классу k через h_{ik}. Тогда, в общем случае, размытая классификация нашего множества $X = \{x_1, ..., x_n\}$ на K классов описывается матрицей $H(X, K) = \{h_{ik}\}$ размерности $N * K$, отражающей принадлежность каждого объекта i к каждому из классов k. Вводятся естественные ограничения на значения элементов матрицы. Принадлежность объекта к любому классу принимает значения от нуля до единицы, а сумма принадлежностей объекта i ко всем классам равна единице: $\sum_{k=1}^{K} h_{ik} = 1$, $0 \leqslant h_{ik} \leqslant 1$. Можно рассматривать эту матрицу как вектор-функцию размерности K от номера объекта, при этом принадлежность объекта i всем классам задается вектор-строкой $H_i = \{h_{i1}, ..., h_{iK}\}$ [36].

3. Способ описания класса. Считается, что объекты k-го класса должны хорошо описываться некоторой моделью (эталоном) этого класса [35]. В соответствии с этим вводится в рассмотрение множество возможных эталонов классов T. Между элементами множества объектов и элементами множества эталонов T вводится некоторая мера близости $S(i, t)$, $(i \in X, t \in T, S(i, t) \geqslant 0)$. Таким образом, любой набор из K классов описывается вектором A эталонов размерности K, $A = (a_1, ..., a_K)$, $(a_k \in T)$. Тогда, близость объекта i к классу k определяется его близостью к соответствующему эталону класса k.

4. Критерий качества классификации. Критерий качества классификации в соответствии с методом обобщенного среднего строится следующим образом:

$$F(H, T) = \sum_{k=1}^{K} \sum_{i=1}^{N} S(i, a_k)\varphi(h_{ik})\,. \tag{8}$$

Этот функционал представляет собой суммарную близость всех объектов ко всем классам, представленным их эталонами, с учетом степени принадлежности. Задача состоит в максимизации критерия (10) по вектор-функции $H(X, K) = \{H_i\}$ принадлежности объектов классам и по вектору эталонов классов $A = (a_1, ..., a_K)$, $a_k \in T$. Здесь $\varphi(h_{ik})$ — монотонно возрастающая функция, отображающая отрезок $[0, 1]$ на себя, причем $\varphi(0) = 0$ и $\varphi(1) = 1$. В литературе рассматривались различные примеры функции $\varphi(h_{ik})$ [36, 37, 38]. Выбор этой функции и ограничения, накладываемые на функцию принадлежности объекта к классу h_{ik} , определяет конкретный тип размытости классификации [36]. Для классификации с фоновым классом, фоновому классу присваивается значение $k = 0$, соответственно функция h_{i0} описывает принадлежность объекта i к фоновому классу.

Четкая классификация

$0 \leqslant h_{ik} \leqslant 1$, $k = 0, ..., K$; $h_{i0} + \sum_{k=1}^{K} h_{ik} = 1$.

Размытая классификация

$0 \leqslant h_{ik} \leqslant 1$, $k = 0, ..., K$; $(h_{i0})^{\lambda} + \sum_{k=1}^{K} (h_{ik})^{\lambda} = 1$, $\lambda > 1$.

В данном случае каждый объект i в оптимальной классификации принадлежит с ненулевым весом ко всем классам, в том числе и к фоновому. Причем мера его принадлежности к фоновому классу тем больше, чем «дальше» объект от нефоновых классов.

Классификация с размытой границей

$0 \leqslant h_{ik} \leqslant 1,\ k = 0, ..., K;\ \sum\limits_{k=0}^{K}(b - h_{ik})^2 = (K-1)b^2 + (b-1)^2$, где b — коэффициент размытости границы. Этот случай является промежуточным между двумя предыдущими случаями: оптимальная классификация выделяет области однозначного отнесения к одному из классов (как к обычному, так и к фоновому), а между ними оказываются зоны неоднозначного отнесения, т.е. размываются только границы классов.

Размытая классификация с четким фоновым классом

$$\begin{cases} h_{i0} = 1;\ h_{ik} = 0;\ k = 1, \ldots, K\,; \\ h_{i0} = 0;\ 0 \leqslant h_{ik} \leqslant 1;\ k = 1, \ldots, K;\ \sum\limits_{k=1}^{K}(h_{ik})^{\lambda} = 1\,. \end{cases}$$

Использование такого ограничения приводит к тому, что фоновый класс — четкий, а разбиение на обычные классы — размытое.

Классификация с размытыми границами между обычными классами и четким фоновым классом

$$\begin{cases} h_{i0} = 1;\ h_{ik} = 0;\ k = 1, \ldots, K\,; \\ h_{i0} = 0;\ 0 \leqslant h_{ik} \leqslant 1;\ k = 1, \ldots, K;\ \sum\limits_{k=1}^{K}(b - h_{ik})^2 = (K-1)b^2 + (b-1)^2\,, \end{cases}$$

где b — коэффициент размытости.

Классификация с четкими обычными классами и размытым фоном

Для того, чтобы размытость была только между фоном и обычными классами, а между классами были четкие границы, нужно ввести единую функцию принадлежности ко всем обычным классам $\hat{h}_i = \sum\limits_{k=1}^{K} h_{ik}$ и ограничения накладывать на $\hat{h}_i$ и h_{i0}, как на функции принадлежности для классификации на два класса:

$0 \leqslant h_{i0} \leqslant 1,\ 0 \leqslant \hat{h}_i \leqslant 1,\ (h_{i0})^{\lambda} + (\hat{h}_i)^{\lambda}$.

Тогда размытость будет только между фоновым классом и объединенным классом, а внутри объединенного класса объект будет относиться к тому классу, к эталону которого он ближе.

Классификация с размытой границей между обычными классами и фоновым классом

Для нашей задачи, когда каждый объект можно, исходя из его физических и биологических свойств, отнести одновременно к нескольким классам, интересно воспользоваться классификацией с разными типами размытости. В работе [36] доказана теоретическая сходимость алгоритма при всех вариантах конкретных функций. Для начала мы исследовали размытую классификацию на разное число классов и со значением показателя размытости $\lambda = 2$ (т. е. фактически размытый вариант кластер-анализа k-средних).

Результаты применения вариационного подхода к кластеризации аминокислотных остатков

В результате применения кластер-анализа аминокислот по геометрическим признакам контактов аминокислот с нуклеотидами в белок-нуклеиновых комплексах (статистике

контактов и площадям контактов, вычисленных с помощью разбиения Вороного–Делоне) были получены следующие основные результаты. Для удобства описания результатов размытой классификации мы будем говорить об отнесении аминокислоты к некоторому классу, если значение ее функции принадлежности к этому классу значительно превышает ее принадлежность к другим классам. Введем в качестве меры отличия размытой и четкой классификации сумму модулей разности принадлежностей, нормированную на число классов и число классифицируемых элементов. В табл. 2 приведены результаты размытой и четкой классификации аминокислотных остатков на 2 класса по признакам контактов и площадей контактов с нуклео-тидами ДНК. В отдельный класс попали аминокислотные остатки ARG и LYS (класс 1). Положительно заряженные аминокислотные остатки аргинин и лизин играют ключевую роль во взаимодействиях с отрицательно заряженной ДНК. Эти остатки могут формировать контакты сразу с несколькими нуклеотидами одновременно. Также эти остатки ответственны за сближение и посадку белков на ДНК [37]. Второй класс образован из остальных 18 аминокислот, и объединяет алифатические аминокислоты, серосодержащие аминокислоты, отрицательно заряженные и слабо заряженный положительно гистидин. Как известно, четкая классификация не позволяет учесть многообразие свойств и их проявлений в тех или иных типах контактов. В результатах размытой классификации видно, что для серина и треонина принадлежность к обоим классам практически одинакова, и отнесение их к какому-то одному классу, как требует четкая классификация, весьма условно. Эти остатки, обладающие гидроксильной группой в боковом радикале, участвуют в образовании водородных связей с нуклеотидами ДНК. Отличие для размытой и четкой классификации по признакам контактов составило 0,1805, по признакам площадей 0,149.

В табл. 3 и 4 приведены результаты размытой и четкой классификации аминокислотных остатков на 4 и 6 классов соответственно, по признакам контактов и площадей контактов с нуклеотидами ДНК. В табл. 3 положительно заряженные аминокислоты аргинин и лизин по-прежнему образуют отдельный класс (3 класс). В то же время результаты размытой классификации указывают на многообразие свойств лизина, входящего с принадлежностью не менее 0,15 во все классы. В отдельный класс объединяются аминокислоты, образующие водородные связи с ДНК: аспарагин, глутамин, глицин, серин, треонин (класс 1). Размытая классификация объединяет гидрофобные остатки Ile, Val, Leu, Ala, а также His, Gln и Tyr в один класс (класс 4). Класс № 2 включает в себя отрицательно заряженные аминокислоты Asp и Glu, серусодержащие аминокислоты метионин и цистеин, а также фенилаланин, пролин и триптофан. Четкая классификация не полностью воспроизводит результаты размытой классификации. Мера отличия составляет 0,238. Так, можно увидеть, что классы 1 и 3 совпадают в обеих классификациях, а в классах 2 и 4 наблюдаются различия. Также, задав некий порог отсечения, можно включать одни и те же аминокислотные остатки одновременно в два и более класса. Результаты классификации по контактам и суммарным площадям контактов также немного различаются между собой. Мера отличия четкой и размытой классификаций по признакам площадей контактов составила 0,225.

В табл. 4, по результатам размытой классификации контактов между аминокислотными остатками и нуклеотидами положительно заряженные аминокислоты аргинин и лизин по-прежнему образуют отдельный класс (1 класс). Аминокислоты, участвующие в образовании водородных связей, оказались рассредоточены по классам 2, 3, 5. Отрицательно заряженные аспарагиновая и глютаминовая кислоты попали в класс с гидрофобными аминокислотами (класс 4). Отдельный класс образовали достаточно редкие ами-

Таблица 2. Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 2 и коэффициенте размытости $\lambda = 2$ (методом k-средних)

	Числа контактов				Площади контактов			
	Размытая кластеризация		Четкая кластеризация		Размытая кластеризация		Четкая кластеризация	
№ класса	1	2	1	2	1	2	1	2
ALA	0,09	0,91	0	1	0,04	0,96	0	1
ARG	0,66	0,34	1	0	0,66	0,34	1	0
ASN	0,28	0,72	0	1	0,22	0,78	0	1
ASP	0,13	0,87	0	1	0,14	0,86	0	1
CYS	0,22	0,78	0	1	0,20	0,80	0	1
GLN	0,15	0,85	0	1	0,14	0,86	0	1
GLU	0,07	0,93	0	1	0,10	0,90	0	1
GLY	0,35	0,65	0	1	0,15	0,85	0	1
HIS	0,07	0,93	0	1	0,05	0,95	0	1
ILE	0,03	0,97	0	1	0,04	0,96	0	1
LEU	0,05	0,95	0	1	0,06	0,94	0	1
LYS	0,82	0,18	1	0	0,95	0,05	1	0
MET	0,18	0,82	0	1	0,14	0,86	0	1
PHE	0,10	0,90	0	1	0,03	0,97	0	1
PRO	0,07	0,93	0	1	0,10	0,90	0	1
SER	0,48	0,52	0	1	0,37	0,63	0	1
THR	0,45	0,55	0	1	0,37	0,63	0	1
TRP	0,20	0,80	0	1	0,15	0,85	0	1
TYR	0,13	0,87	0	1	0,23	0,77	0	1
VAL	0,04	0,96	0	1	0,06	0,94	0	1

нокислоты цистеин, триптофан и метионин (класс 6). Здесь также результаты размытой классификации отличаются от результатов четкой классификации. Результаты классификации суммарных площадей в целом повторяют результаты классификации контактов, с некоторыми отличиями. Отличие для размытой и четкой классификации по признакам контактов составило 0,169, по признакам площадей 0,231. Преимущество размытой классификации наглядно видно на примере лизина (см. табл. 4). Видно, что лизин входит во все классы с принадлежностью не менее 0,1. В действительности, лизин участвует во всех возможных взаимодействиях с ДНК – образовании ионных мостиков, водородных связей, ван дер ваальсовых взаимодействий. Таким образом, размытая классификация позволяет учесть многообразие свойств и проявлений этих свойств аминокислот. Интерпретация результатов классификации зачастую представляет самостоятельную задачу, поскольку только базовых свойств аминокислот насчитывается более десяти, всего же на данный момент в базе данных AAindex содержится 544 различных свойств для каждого типа аминокислотного остатка [39].

Размытая классификация при увеличении числа классов более 6 создает дублирующиеся классы, с одинаковым составом, что указывает на нецелесообразность дальнейшего разделения. Тем самым позволяет определить естественное максимальное число классов.

Таблица 3. Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 4 и коэффициенте размытости $\lambda = 2$

	Числа контактов								Площади контактов							
	Размытая кластеризация				Четкая кластеризация				Размытая кластеризация				Четкая кластеризация			
№ класса	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
ALA	0,19	0,27	0,03	0,51	0	0	0	1	0,08	0,28	0,01	0,63	0	0	0	1
ARG	0,12	0,09	0,69	0,10	0	0	1	0	0,05	0,04	0,86	0,05	0	0	1	0
ASN	0,61	0,15	0,04	0,20	1	0	0	0	0,59	0,16	0,03	0,22	1	0	0	0
ASP	0,10	0,55	0,03	0,33	0	0	0	1	0,11	0,57	0,03	0,29	0	1	0	0
CYS	0,17	0,43	0,06	0,33	0	1	0	0	0,17	0,46	0,05	0,33	0	1	0	0
GLN	0,28	0,27	0,04	0,41	0	0	0	1	0,31	0,26	0,03	0,40	1	0	0	0
GLU	0,07	0,48	0,02	0,43	0	0	0	1	0,06	0,69	0,01	0,23	0	1	0	0
GLY	0,83	0,07	0,02	0,08	1	0	0	0	0,37	0,23	0,03	0,36	1	0	0	0
HIS	0,08	0,41	0,02	0,50	0	0	0	1	0,09	0,31	0,01	0,59	0	0	0	1
ILE	0,04	0,21	0,01	0,74	0	0	0	1	0,07	0,35	0,01	0,56	0	0	0	1
LEU	0,06	0,40	0,01	0,53	0	0	0	1	0,07	0,49	0,01	0,43	0	0	0	1
LYS	0,23	0,15	0,46	0,16	0	0	1	0	0,32	0,21	0,25	0,22	0	0	1	0
MET	0,13	0,50	0,04	0,33	0	1	0	0	0,11	0,56	0,03	0,30	0	1	0	0
PHE	0,04	0,78	0,01	0,17	0	0	0	1	0,06	0,17	0,01	0,76	0	0	0	1
PRO	0,06	0,55	0,01	0,37	0	0	0	1	0,07	0,65	0,01	0,26	0	0	0	1
SER	0,65	0,13	0,06	0,16	1	0	0	0	0,71	0,11	0,03	0,14	1	0	0	0
THR	0,71	0,11	0,05	0,13	1	0	0	0	0,69	0,12	0,04	0,15	1	0	0	0
TRP	0,15	0,46	0,05	0,33	0	1	0	0	0,12	0,55	0,03	0,30	0	1	0	0
TYR	0,25	0,27	0,04	0,44	0	0	0	1	0,61	0,15	0,03	0,21	1	0	0	0
VAL	0,07	0,22	0,01	0,70	0	0	0	1	0,09	0,37	0,02	0,53	0	0	0	1

Заключение

Для широкого круга биоинформатических исследований представляет большой интерес уменьшение сложности описания 20 стандартных аминокислот путем их разбиения на группы и создания так называемого «вырожденного алфавита». Хотя не существует универсального способа классификации аминокислот, имеются многочисленные примеры использования различных методов и алгоритмов кластер-анализа, с одной стороны и различных типов исходной информации для такой группировки (физико-химические свойства, мутации, эволюционные замены и т. д.), с другой стороны. Впервые в данной работе в качестве исходной информации для классификации аминокислот используются данные о пространственных контактах между аминокислотными остатками и нуклеотидами в структурах комплексов белок-ДНК. При этом для определения таких контактов применяется метод пространственного разбиения Вороного–Делоне. Кроме того, впервые учитывается площадь контакта между соседними атомами. При помощи математической модели показан неслучайный характер таких контактов, а именно около 30% всех контактов между аминокислотами и нуклеотидами в комплексах белок-ДНК являются неслучайными. На основе классических методов кластер-анализа (иерархических, типа k-средних, и других) и с применением различных мер близости построены классификации аминокис-

Таблица 4. Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 6 и коэффициенте размытости $\lambda = 2$

	Числа контактов												Площади контактов											
	Размытая кластеризация						Четкая кластеризация						Размытая кластеризация						Четкая кластеризация					
№ класса	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
ALA	0,01	0,79	0,05	0,08	0,03	0,03	0	1	0	0	0	0	0,01	0,50	0,10	0,16	0,06	0,17	0	1	0	0	0	0
ARG	0,84	0,03	0,04	0,03	0,04	0,03	1	0	0	0	0	0	0,92	0,01	0,02	0,02	0,02	0,01	1	0	0	0	0	0
ASN	0,02	0,14	0,58	0,08	0,13	0,05	0	0	1	0	0	0	0,01	0,06	0,61	0,17	0,10	0,04	0	0	0	1	0	0
ASP	0,02	0,18	0,10	0,34	0,07	0,28	0	0	0	0	1	0	0,01	0,20	0,07	0,10	0,05	0,57	0	0	0	0	0	1
CYS	0,03	0,13	0,09	0,19	0,07	0,49	0	0	0	0	0	1	0,03	0,23	0,12	0,14	0,09	0,39	0	0	0	0	0	1
GLN	0,02	0,49	0,17	0,15	0,09	0,07	0	1	0	0	0	0	0,01	0,12	0,20	0,52	0,08	0,07	0	0	0	1	0	0
GLU	0,02	0,20	0,08	0,50	0,06	0,14	0	0	0	0	1	0	0,01	0,27	0,07	0,10	0,05	0,50	0	0	0	0	1	0
GLY	0,01	0,09	0,65	0,05	0,16	0,04	0	0	1	0	0	0	0,01	0,09	0,24	0,53	0,08	0,06	0	0	0	1	0	0
HIS	0,01	0,17	0,07	0,57	0,05	0,12	0	0	0	1	0	0	0,01	0,56	0,09	0,13	0,06	0,16	0	0	1	0	0	0
ILE	0,01	0,11	0,04	0,75	0,03	0,06	0	0	0	1	0	0	0,01	0,66	0,06	0,09	0,04	0,13	0	0	1	0	0	0
LEU	0,01	0,08	0,03	0,81	0,02	0,05	0	0	0	1	0	0	0,01	0,64	0,06	0,09	0,04	0,17	0	0	0	0	1	0
LYS	0,25	0,14	0,17	0,12	0,20	0,11	1	0	0	0	0	0	0,15	0,14	0,19	0,17	0,23	0,13	1	0	0	0	0	0
MET	0,01	0,05	0,03	0,08	0,02	0,82	0	0	0	0	0	1	0,01	0,19	0,07	0,09	0,05	0,59	0	0	0	0	0	1
PHE	0,02	0,15	0,08	0,47	0,06	0,22	0	0	0	0	1	0	0,01	0,52	0,10	0,16	0,06	0,14	0	1	0	0	0	0
PRO	0,01	0,13	0,06	0,63	0,04	0,12	0	0	0	1	0	0	0,01	0,34	0,08	0,11	0,05	0,41	0	0	0	0	1	0
SER	0,01	0,05	0,12	0,04	0,76	0,03	0	0	1	0	0	0	0,01	0,05	0,13	0,08	0,70	0,04	0	0	0	1	0	0
THR	0,02	0,06	0,16	0,05	0,68	0,03	0	0	1	0	0	0	0,01	0,05	0,12	0,08	0,71	0,04	0	0	0	1	0	0
TRP	0,02	0,10	0,06	0,15	0,05	0,63	0	0	0	0	0	1	0,02	0,19	0,08	0,10	0,06	0,56	0	0	0	0	0	1
TYR	0,01	0,59	0,12	0,14	0,07	0,06	0	1	0	0	0	0	0,01	0,08	0,53	0,19	0,13	0,05	0	0	0	1	0	0
VAL	0,01	0,27	0,09	0,47	0,06	0,10	0	0	0	1	0	0	0,01	0,52	0,09	0,14	0,06	0,18	0	0	1	0	0	0

лотных остатков и проанализированы их свойства и выявлены инварианты кластеризации аминокислот. В некоторых случаях объединение аминокислот в классы по признакам пространственных контактов с нуклеотидами совпадает с результатами кластеризации на основе физико-химических свойств аминокислот. Это является еще одним подтверждением адекватности предлагаемого подхода. Было показано совпадение результатов классификаций для выборки в целом и двух ее подвыборок. В то же время единое жесткое разбиение аминокислот на фиксированные группы не может отразить сложный характер взаимодействия аминокислот белка и нуклеотидов ДНК, существующий в природе. В связи с этим предложено использовать вариационные методы для построения различных типов размытой классификации аминокислот (размытая классификация, классификация с перекрывающимися классами, классификация с размытыми границами и с фоновым классом), позволяющие учесть разные аспекты взаимодействий ДНК-белок. Показано, что применение размытой классификации позволяет более адекватно описывать разные аспекты белок-нуклеинового взаимодействия.

Литература

[1] *Gurskii G. V., Tumanian V. G., Zasedatelev A. S., Zhuze A. L., Grokhovskii S. L., Gottikh B. P.* A code governing specific binding of regulatory proteins to DNA and structure of stereospecific sites of regulatory proteins // *Mol. Biol. Mosk.*, 1975. Vol. 9, No. 5. P. 635–651.

[2] *Gurskii G. V., Zasedatelev A. S.* Precise relationships for calculating the binding of regulatory proteins and other lattice ligands in double-stranded polynucleotides // *Biofizika*, 1978. Vol. 23, No. 5. P. 932–946.

[3] *Jordan S. R., Pabo C. O.* Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions // *Science*, 1988. Vol. 242, No. 4880. P. 893–899.

[4] *Brennan R. G., Roderick S. L., Takeda Y., Matthews B. W.* Protein-DNA conformational changes in the crystal structure of a lambda Cro-operator complex // *Proc. Natl. Acad. Sci. U.S.A.*, 1990. Vol. 87, No. 20. P. 8165–8169.

[5] *Schultz S. C., Shields G. C., Steitz T. A.* Crystal structure of a CAP-DNA complex: The DNA is bent by 90 degrees // *Science*, 1991. Vol. 253, No. 5023. P. 1001–1007.

[6] *Bohm H. J.* The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure // *J. Comput. Aided Mol. Des.*, 1994. Vol. 8, No. 3. P. 243–256.

[7] *Aqvist J., Fothergill M.* Computer simulation of the triosephosphate isomerase catalyzed reaction // *J. Biol. Chem.*, 1996. Vol. 271, No. 17. P. 10010–10016.

[8] *Eldridge M. D., Murray C. W., Auton T. R., Paolini G. V., Mee R. P.* Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes // *J. Comput. Aided Mol. Des.*, 1997. Vol. 11, No. 5. P. 425–445.

[9] *Cozzini P., Fornabaio M., Marabotti A., Abraham D. J., Kellogg G. E., Mozzarelli A.* Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water // *J. Med. Chem.*, 2002. Vol. 45, No. 12. P. 2469–2483.

[10] *Lesser D. R., Kurpiewski M. R., Jen-Jacobson L.* The energetic basis of specificity in the Eco RI endonuclease–DNA interaction // *Science*, 1990. Vol. 250, No. 4982. P. 776–786.

[11] *Draper D. E.* Protein–DNA complexes: The cost of recognition // *Proc. Natl. Acad. Sci. U.S.A.*, 1993. Vol. 90, No. 16. P. 7429–7430.

[12] *Mandel-Gutfreund Y., Margalit H.* Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites // *Nucleic Acids Res.*, 1998. Vol. 26, No. 10. P. 2306–2312.

[13] *Mandel-Gutfreund Y., Schueler O., Margalit H.* Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: In search of common principles // *J. Mol. Biol.*, 1995. Vol. 253, No. 2. P. 370–382.

[14] *Choo Y., Klug A.* Physical basis of a protein-DNA recognition code // *Curr. Opin. Struct. Biol.*, 1997. Vol. 7, No. 1. P. 117–125.

[15] *Jones S., van Heyningen P., Berman H. M., Thornton J. M.* Protein-DNA interactions: A structural analysis // *J. Mol. Biol.*, 1999. Vol. 287, No. 5. P. 877–896.

[16] *Oda M., Nakamura H.* Thermodynamic and kinetic analyses for understanding sequence-specific DNA recognition // *Genes Cells*, 2000. Vol. 5, No. 5. P. 319–326.

[17] *Pabo C. O., Nekludova L.* Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? // *J. Mol. Biol.*, 2000. Vol. 301, No. 3. P. 597–624.

[18] *Benos P. V., Lapedes A. S., Stormo G. D.* Is there a code for protein–DNA recognition? Probab(ilistical)ly // *Bioessays*, 2002. Vol. 24, No. 5. P. 466–475.

[19] *Luscombe N. M., Thornton J. M.* Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity // *J. Mol. Biol.*, 2002. Vol. 320, No. 5. P. 991–1009.

[20] *Benos P. V., Lapedes A. S., Stormo G. D.* Probabilistic code for DNA recognition by proteins of the EGR family // *J. Mol. Biol.*, 2002. Vol. 323, No. 4. P. 701–727.

[21] *Gorfe A. A., Jelesarov I.* Energetics of sequence-specific protein–DNA association: Computational analysis of integrase Tn916 binding to its target DNA // *Biochemistry*, 2003. Vol. 42, No. 40. P. 11568–11576.

[22] *Venkatarajan M. S., Braun W.* New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties // *J. Mol. Model.*, 2001. Vol. 7, No. 12. P. 445–453.

[23] *Shen B., Bai J., Vihinen M.* Physicochemical feature-based classification of amino acid mutations // *Protein Eng. Des. Sel.*, 2008. Vol. 21, No. 1. P. 37–44.

[24] *Kosiol C., Goldman N., Buttimore N. H.* A new criterion and method for amino acid classification // *J. Theor. Biol.*, 2004. Vol. 228, No. 1. P. 97–106.

[25] *Rogov S. I., Nekrasov A. N.* A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences // *Protein Eng.*, 2001. Vol. 14, No. 7. P. 459–463.

[26] *May A. C.* Towards more meaningful hierarchical classification of amino acid scoring matrices // *Protein Eng.*, 1999. Vol. 12, No. 9. P. 707–712.

[27] *Davies M. N., Secker A., Halling-Brown M., Moss D. S., Freitas A. A., Timmis J., Clark E., Flower D. R.* GPCRTree: Online hierarchical classification of GPCR function // *BMC Res. Notes*, 2008. Vol. 1. P. 67.

[28] *Davies M. N., Secker A., Freitas A. A., Clark E., Timmis J., Flower D. R.* Optimizing amino acid groupings for GPCR classification // *Bioinformatics*, 2009. Vol. 11, No. 1. P. 111–122.

[29] *Anashkina A., Kuznetsov E., Esipova N., Tumanyan V.* Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces // *Proteins*, 2007. Vol. 67, No. 4. P. 1060–1077.

[30] *Anashkina A. A., Tumanyan V. G., Kuznetsov E. N., Galkin A. V., Esipova N. G.* Geometrical analysis of protein-DNA interactions on the basis of the Voronoi-Delaune tessellation // *Biofizika*, 2008. Vol. 53, No. 3. P. 402–406. (in Russ.)

[31] *Medvedev N. N.* Voronoi-Delaunay Method in Noncrystal Systems Investigations. Novosibirsk: SO RAS, 2000. (in Russ.)

[32] *Raushenbakh G. V.* Proximity and similarity measures // in Non-numerical information analysis in social science. M.: Nauka, 1985. P. 169–203. (in Russ.)

[33] *Mirkin B. G.* Cluster Analysis for Decision Making: Review. M.: HSE, 2011. (in Russ.)

[34] *Gunasekaran K., Ramakrishnan C., Balaram P.* Disallowed Ramachandran conformations of amino acid residues in protein structures // *J. Mol. Biol.*, 1996. Vol. 264, No. 1. P. 191–198.

[35] *Diday E.* Data analysis methods. (Trans.from fr. Diday E. et collaborateurs. Optimisation en classification automatique. Paris: Institut national de rechercher en informatique et en automatique, 1979) Moscow: Finansy i Statistika, 1985. 357 p. (in Russ.)

[36] *Bauman E. V., Bludjan N. O.* Metody nahozhdenija global'nyh jekstremumov funkcionalov v zadache klassifikacionnogo analiza dannyh // *Trudy Instituta problem upravlenija RAN*, 2001. Vol. XIII. P. 129–136. (in Russ.)

[37] *Zadeh L. A.* Fuzzy sets as a basis for a theory of possibility // *Fuzzy Sets Systems*, 1978. Vol. 1. P. 3–28.

[38] *Bezdek J. C.* A convergence theorem for the fuzzy ISODATA clusters algorithms // *IEEE Trans. Pattern Analysis Machine Intelligence*, 1980. P. 1–8.

[39] *Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.* AAindex: Amino acid index database, progress report 2008 // *Nucleic Acids Res.*, 2008. Vol. 36. Database issue. P. D202–D205.

Интеллектуальные возможности гипертрейс-преобразования: конструирование признаков с заданными свойствами*

***Н. Г. Федотов*[1], *А. А. Семов*[1], *А. В. Моисеев*[2]**

fedotov@pnzgu.ru, mathematik_aleksey@mail.ru, moigus@mail.ru

Россия, г. Пенза [1]Пензенский государственный университет, ул. Красная, 40; [2]Пензенский государственный технологический университет, проезд Байдукова/ул. Гагарина, 1 а/11

В настоящей статье предлагается новый подход к распознаванию трехмерных (3D) объектов, основанный на современных методах стохастической геометрии и функционального анализа. Данный метод обладает рядом преимуществ и возможностями интеллектуального анализа данных. Так, признаки имеют гипертриплетную композиционную структуру, которая способствует не только легкой машинной реализации этого алгоритма, но и конструированию большого числа признаков. Благодаря построению строгой математической модели, аналитик может строить признаки не интуитивно, а аналитически, описывая каждый класс объектов и их особенности (в частности, конструирование геометрических признаков). Трехмерное трейс преобразование позволяет создавать инвариантное описание пространственного объекта, которое является более устойчивым к искажениям и координатным шумам, чем описание, получаемое в результате процедуры нормализации объекта. Возможность регулировать свойства построенных признаков заметно повышает интеллектуальные возможности 3D трейс преобразования, что, несомненно, является его преимуществом. Доказательством разработанной теории и математической модели является множество построенных теоретических примеров гипертриплетных признаков, имеющихся описанные определенные свойства. В статье анализируется роль функционалов, входящих в композиционную структуру гипертриплетного признака. Описываются расширенные возможности 3D трейс преобразования, в частности, извлечение в той же технике сканирования информации о пространственном положении и ориентации трехмерного объекта. Приводится описание многих способов интеллектуального анализа 3D изображений. Например, одной из интеллектуальных способностей предлагаемого метода является высокоуровневая предобработка, обработка и постобработка 3D изображения в одной технике сканирования.

Ключевые слова: гипертрейс-преобразование; 3D распознавание образов; интеллектуальный анализ 3D изображений; инвариантность и сенситивность признаков; определение параметров пространственного объекта

Intelligent capabilities hypertrace transform: Constructing features with predetermined properties*

***N. G. Fedotov*[1], *A. A. Syemov*[1], *A. V. Moiseev*[2]**

[1]Penza State University, Penza, Russia; [2]Penza State Technological University, Penza, Russia

Background: In recent decades, the emphasis in the analysis and pattern recognition shifts from two-dimensional (2D) to three-dimensional (3D) images, because 3D design allows to use more information about the object. Three-dimensional modeling gives possibility to see object from different angles, in particular, allows to analyze its spatial form.

*Работа выполнена при финансовой поддержке РФФИ, проект № 12-07-00501.

Methods: In this article, a new approach to the 3D objects' recognition based on modern methods of stochastic geometry and functional analysis is proposed. This method has many advantages and data mining capabilities. Thus, features have hypertriplet composite structure, which provide not only easy machine implementation of this algorithm, but construction of a large number of features. Due to building a rigorous mathematical model, the analyst can construct analytical and not intuitive features, describing each object class and their features (in particular, constructing geometric features).

Results: Three-dimensional trace transform allows to create invariant description of spatial object, which is more resistant to distortion and coordinate noise than the description obtained as a result of the object normalization procedure. Possibility of regulating constructed features' properties significantly increases intellectual capabilities of 3D trace transform that is undoubtedly its advantage. Proof developed theory and the mathematical model is variety constructed theoretical examples of hypertriplet features having described particular properties.

Concluding Remarks: In the article, the role of functional included in composite structure of hypertriplet feature is analyzed. Extended possibilities of 3D trace transform, in particular, extracting in the same scanning technique the information about the spatial position and orientation of 3D object, are described. Description of many ways of 3D image mining is proposed. For example, one of the intellectual abilities of the proposed method is a high-level preprocessing, processing, and postprocessing of 3D images in one scanning technique.

Keywords: hypertrace transform; 3D pattern recognition; 3D image mining; invariance and sensitiveness features; defining the spatial object parameters

Введение

В последние десятилетия акцент в анализе и распознавания образов смещается с двухмерных (2D) на 3D изображения, так как трехмерное моделирование и конструирование позволяют полнее учитывать информацию об объекте, дает возможность видеть его с разных углов обзора и, в частности, позволяет анализировать его пространственную форму [1].

Интеллектуальная компьютерная обработка как 2D, так и 3D изображений находится пока не на высоком уровне, несмотря на возрастающую потребность в данных методах [2]. Это обусловлено отнюдь не низкой мощностью вычислительных средств, а недостаточно развитыми теоретическими подходами. Самостоятельное принятие решение и искусственный интеллект роботов будет невозможен или крайне неэффективен при слабом уровне машинного зрения. Так, транспортные роботы при плохом качестве машинного зрения не смогут обеспечить надежную ориентацию и движение в пространстве.

С развитием и усложнением технических устройств актуальной становится также задача анализа особенностей 3D объектов, вычисление различных их метрических свойств и признаков. Так, например, для ориентирования роботизированного мобильного комплекса на неизвестной местности актуальной является задача анализа ситуации с помощью бортового вычислителя, исходя из информации, поставляемой системой технического зрения в реальном времени. Необходимо оперативно не только распознавать 3D объекты, но и определять скорости этих объектов и расстояния до них, определять относительно мировой и бортовой систем координат их положение и ориентацию в пространстве и т.п.

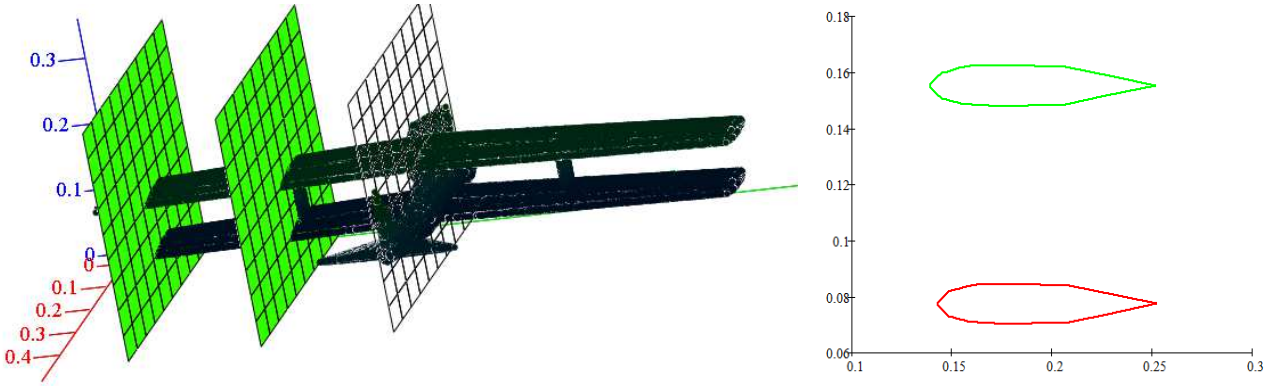
В данной статье предлагается новый подход к конструированию признаков 3D изображения, дающие инвариантное описание объекта при любой его пространственной ориентации – гипертрейс-преобразование (или 3D трейс-преобразование), обладающего не только

множеством дополнительных возможностей анализа 3D объектов, но и высокими интеллектуальными способностями.

Математическая модель гипертрейс-преобразования

Пусть F — исходная трехмерная модель. Определим плоскость $B(\eta, r) = \{x | x^T \eta = r\}$ как касательную к сфере с центром в начале координат и с радиусом r в точке (η, r), где $\eta = [\cos\varphi \sin\omega, \sin\varphi \cdot \sin\omega, \cos\omega]$ — единичный вектор в R^3, r, ω и φ — сферические координаты.

Сканирование исходного пространственного объекта F осуществляется сеткой параллельных плоскостей, определяемой парой углов (ω, φ) с расстоянием Δr между плоскостями. Взаимное положение 3D объекта F и каждой сканирующей плоскости $B(\eta(\omega, \varphi), r)$ характеризуется числом G по некоторому правилу $\mathrm{Hyper}\, T : G = \mathrm{Hyper}\, T\,(F \bigcap B\,(\eta(\omega, \varphi), r))$. В качестве указанной характеристики могут выступать число пересечений плоскости с исходным объектом, площадь сечения или свойства окрестности такого сечения и т. п. (рис. 1).



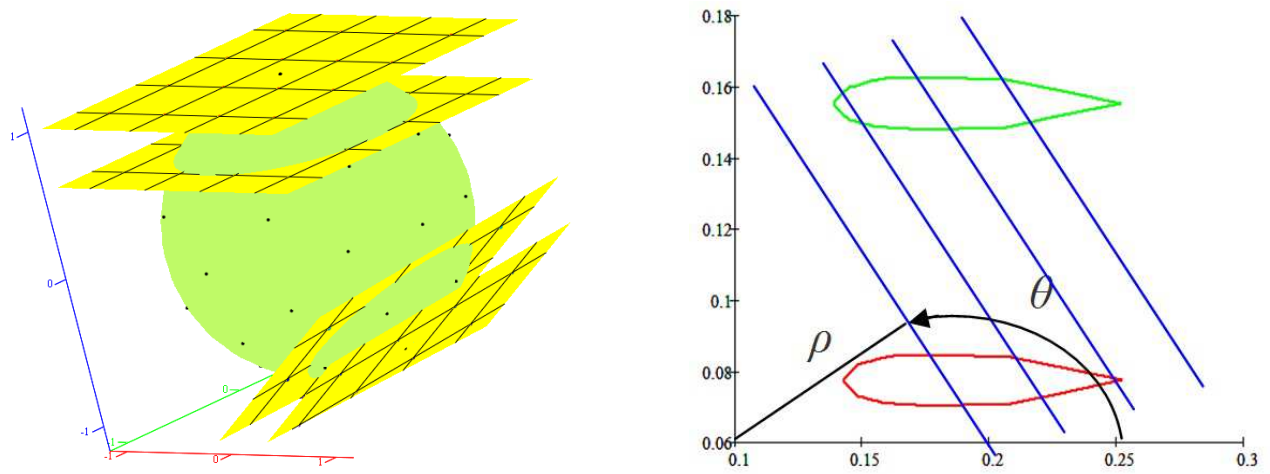
(а) Процесс сканирования 3D объекта плоскостями (б) Одно из сечений 3D объекта

Рис. 1. Особенности сканирования 3D объекта

Затем сканирование производится для новой пары значений углов $(\omega + \Delta\omega, \varphi + \Delta\varphi)$, получивших дискретные приращения $\Delta\varphi$, сеткой параллельных плоскостей с тем же шагом Δr. К пересечению новой плоскости $B(\eta(\omega + \Delta\omega, \varphi + \Delta\varphi), r)$ и 3D объекта F вновь применяется ранее выбранное правило $HyperT$. Важно отметить, что изменять углы важно не произвольным образом [3], а согласно построению опорной сетки, чтобы плотность плоскостей в пространстве была равномерной (рис. 2).

После перебора различных вариантов пар углов (ω, φ) получаем множество чисел G, которое формирует гипертрейс-матрицу 3ТМ, у которой ось $O\omega$ будет направлена горизонтально, ось $O\varphi$ — вертикально, ось Or — вглубь. После формирования матрицы $3TM$ с помощью гиперфункционала $\mathrm{Hyper}\, P$ (например, максимальный элемент строки) обрабатываются ее глубинные строки. В результате данная матрица станет двумерной.

Таким образом, признак 3D изображения получается после обработки строк и столбцов матрицы 3ТМ гиперфункционалами $\mathrm{Hyper}\,\Theta$, $\mathrm{Hyper}\,\Omega$ и $\mathrm{Hyper}\, P$. Получается структура гипертриплетных признаков в виде композиции функционалов — последовательное применение указанных выше гиперфункционалов, каждый из которых сокращает размерность матрицы 3ТМ на единицу [4]:



(а) Формирование опорной равномерной сетки (б) Сканирование сечения сеткой параллельных прямых

Рис. 2. Формирование гипертрейс-матрицы 3ТМ 3D объекта

$$\mathrm{Res}(F) = \mathrm{Hyper}\,\Theta \circ \mathrm{Hyper}\,\Omega \circ \mathrm{Hyper}\,P \circ \mathrm{Hyper}\,T(F_{sect}).$$

Сканирование получаемых в сечение фигур F_{sect} осуществляется сеткой параллельных прямых $l(\rho, \theta)$ с расстоянием $\Delta\rho$ между линиями, где ρ, θ — полярные координаты прямой в плоскости сечения. Взаимное положение изображения F_{sect} и каждой сканирующей линии $l(\rho, \theta)$ характеризуется числом, вычисляемым по некоторому правилу $T : g = T(F \bigcap l)$ (рис. 2).

Затем сканирование производится для нового значения угла $\theta + \Delta\theta$, получившего дискретное приращение $\Delta\theta$, сеткой параллельных прямых в той же плоскости сечения F_{sect} и с тем же шагом $\Delta\rho$. К пересечению новой прямой $l(\rho, \theta)$ и сечения F_{sect} применяется ранее выбранное правило T.

Результат вычислений трейс-функционала T зависит от двух параметров прямой ρ и θ, на основе которых формируется трейс-матрица ТМ. Триплетный признак сечения получается после обработки строк и столбцов матрицы ТМ функционалами Θ и P, каждый из которых последовательно сокращает размерность матрицы на единицу (аналогично признаку 3D изображения).

Таким образом, признак 2D фигуры сечения имеет следующую структуру [5]:

$$\Pi(F_{\mathrm{sect}}) = \mathrm{Hyper}\,T(F_{\mathrm{sect}}) = \Theta \circ P \circ T(F_{\mathrm{sect}} \bigcap l(\theta, \rho)).$$

Благодаря композиции функционалов и гиперфункционалов, входящих в структуру признака $\Pi(F_{\mathrm{sect}})$ и $\mathrm{Res}(F)$ соответственно, возможно получение огромного числа признаков. Причем некоторые из них имеют явную геометрическую интерпретацию, что облегчает задачу построения признаков и повышает их различающую силу. Специфичная структура гипертриплетных и триплетных признаков позволяет строить признаки как сенситивные, так и инвариантные к группе движений и масштабированию, что повышает интеллектуальность и гибкость 3D трейс-метода при распознавании объектов.

В заключении стоит отметить, что приставка «гипер-» для всех терминов здесь и далее означает, что речь идет о пространственном объекте в целом и его свойствах, а не о каком-либо его конкретном сечении и их особенностях. Так, гиперфункционал $\mathrm{Hyper}\,P$ по сути

ничем не отличается от функционала P с той лишь разницей, что первый термин для удобства восприятия употребляется при обработке глубинных строк гипертрейс-матрицы 3TM, а второй термин — при обработке столбцов трейс-матрицы TM.

Функционалы, инвариантные к переносу

Свойство инвариантности к переносу для любого функционала ζ будет иметь вид: $\zeta\left(f(x+b)\right) = \zeta\left(f(x)\right)$ для всех допустимых $f(x)$ и $\forall b$.

Так как объект сканируется сеткой параллельных плоскостей, то перемещение исходного 3D объекта изменяет только размер гипертрейс-матрицы на количество нулевых элементов внутри глубинных строк (вдоль осей $0r$). При этом при этом все глубинные строки матрицы останутся на своих местах и значение признака $\Pi(F_{\text{sect}})$ не изменяется. Таким образом, для того чтобы признак $\text{Res}(F)$ (и признак $\Pi(F_{\text{sect}})$ плоского сечения) был инвариантен к переносу, необходимо и достаточно, чтобы гиперфункционал $\text{Hyper}\,P$ (и функционал P) обладали свойством: $\underset{x+b\neq 0}{\zeta}\left(f(x+b)\right) = \underset{x\neq 0}{\zeta}\left(f(x)\right)$ для всех допустимых $f(x)$ и $\forall b$.

Приведем пример признака, инвариантного к операции трансляции и сенситивного к операциям ротации и гомотетии. Этот пример будет объяснен подробно, а все остальные теоретические примеры будут описаны кратко:

$$\text{Res}(F) = \text{Hyper}\,\Theta \circ \text{Hyper}\,\Omega \circ \text{Hyper}\,P \circ \text{Hyper}\,T(\Theta \circ P \circ T),$$

где $T(F_{\text{sect}} \bigcap l(\theta,\rho)) = \max\limits_{t} f(\theta,\rho,t)$, $P\left(g(\theta,\rho)\right) = \min\limits_{i} f(\theta,\rho_i)/\sum_i f(\theta,\rho_i)$, $\Theta(g(\theta)) = \arg_k f(\theta)$, $\text{Hyper}\,T(F \bigcap B(\eta(\omega,\varphi),r)) = \Pi\left(F_{\text{sect}}\right) = G(\omega,\varphi,r)$, $\text{Hyper}\,P\left(\Pi(\omega,\varphi,r)\right) = \max\limits_{r} G(\omega,\varphi,r)/\min\limits_{r} G(\omega,\varphi,r)$, $\text{Hyper}\,\Omega\left(\Pi(\omega,\varphi)\right) = \underset{\omega}{\text{NumMax}}\,G(\omega,\varphi)$, $\text{Hyper}\,\Theta\left(\Pi(\varphi)\right) = \arg_k G(\varphi)$, $f(\theta,\rho,t)$ — длина t-го отрезка, высекаемого ρ_j-й прямой под θ_i-м углом в плоскости сечения F_{sect}, $\Pi\left(F_{\text{sect}}\right) = G(\omega,\varphi,r)$ — признак сечения, получаемого r_h-й плоскости под $(\omega_w,\,\varphi_s)$-м углом, $\arg_k V$ — функционал, вычисляющий k-ю координату вектора V, $\text{NumMax}\,f(x)$ — функционал, вычисляющий число локальных максимумов функции $f(x)$.

Так, для приведенного признака функционал T находит для каждой сканирующей прямой из сетки параллельных прямых под разными углами максимальный отрезок прямой, содержащийся в плоскости сечения F_{sect}. Функционал P для каждой сетки параллельных прямых из всего множества сеток под разными углами θ в плоскости сечения вычисляет отношение минимальной длины среди указанных выше максимальных отрезков ко всей сумме этих отрезков внутри каждой сетки. Функционал Θ выбирает среди подсчитанных выше отношений значение для k-й сетки. Функционал $\text{Hyper}\,T$ соответствует признаку сечения, который вычисляется при помощи композиции функционалов $\Theta \circ P \circ T$. Функционал $\text{Hyper}\,P$ для каждой сетки параллельных плоскостей вычисляет отношение максимального значения признака сечений к минимальному внутри каждой сетки. Функционал $\text{Hyper}\,\Omega$ вычисляет число локальных максимумов функции, образованных дискретным рядом горизонтальных строк (ось $O\omega$), содержащих значения отношений максимума к минимуму. Функционал $\text{Hyper}\,\Theta$ выбирает среди подсчитанных выше значений k-ый элемент (ось $O\varphi$).

Класс признаков, аналогичный приведенному выше, полезен в некоторых частных задачах, например, когда заранее известно, что анализируемые 3D объекты отличаются от эталонных в базе только переносом. Преимуществом данного класса признаков может являться их низкие вычислительные затраты, когда все функционалы, кроме T и $\text{Hyper}\,T$,

заменяются на $\arg_k V$ (ссылка на k-й элемент массива), или же их вычисление вообще игнорируется (сканирование плоскостями и прямыми происходит только под одним углом обзора (ω, φ) и θ соответственно).

Однако в общем случае на практике необходимо строить признаки, инвариантные хотя бы к повороту и переносу 3D изображения и сенситивные к масштабированию. Свойство инвариантности к переносу является очень полезным при решении различных задач, так как благодаря этому свойству не нужно производить нормализацию по переносу и определять центр масс объекта, сокращая, тем самым, дополнительные вычислительные затраты и получая признаки, более устойчивые к искажениям и координатному шуму.

Функционалы, сенситивные к переносу

Свойство сенситивности к переносу для любого функционала ζ будет иметь вид: $\zeta\left(f(x+b)\right) = \zeta\left(f(x)\right) - b$ или $\zeta\left(f(x+b)\right) = \zeta\left(f(x)\right) + b$, или в общем виде $\zeta\left(f(x+b)\right) = \zeta\left(f(x)\right) + bk$ для всех допустимых $f(x)$ и $\forall b$, $\forall k$.

Возможность регулировать свойства конструируемых признаков повышает интеллектуализацию анализа и обработки 3D изображений, так как не только расширяет количество полезных признаков, эффективных для того или иного класса 3D объектов, но и позволяет извлекать дополнительную информацию об объекте, определять параметры преобразования самого объекта (отличия объекта от эталона) и т. п.

Так, например, покажем, как, вычисляя два сенситивных к переносу признака 3D объекта, можно получить в одной технике признак $\mathrm{Res}(F)$, инвариантный к переносу, повороту и масштабированию, а также уравнения граней выпуклого многогранника, содержащего внутри себя исходный объект.

В качестве сенситивного к переносу признака возьмем следующий:

$$\mathrm{Sen}\,(F) = \mathrm{Hyper}\,\Theta \circ \mathrm{Hyper}\,\Omega \circ \mathrm{Hyper}\,P \circ \mathrm{Hyper}\,T(\Theta \circ P \circ T),$$

где $T(\theta, \rho) = \left(\max\limits_t f(\theta, \rho, t) + \min\limits_t f(\theta, \rho, t)\right)/2$; $P(\theta) = \min\limits_i f(\theta, \rho_i)/\sum_i f(\theta, \rho_i)$; $\Theta = \sqrt[k]{\prod_k f(\theta_k)}$, если $|f(\theta_k)| \geqslant 0{,}0001$; $\mathrm{Hyper}\,T(\omega, \varphi, r) = \Pi(F_{\mathrm{sect}}) = G(\omega, \varphi, r)$; $\mathrm{Hyper}\,P(\omega, \varphi) = \arg_{\mathrm{first}} G(\omega, \varphi, r)$; $\mathrm{Hyper}\,\Omega(\varphi) = \max\limits_\omega G(\omega, \varphi)$; $\mathrm{Hyper}\,\Theta = \min\limits_\varphi G(\varphi)$; $\arg_{\mathrm{first}}(\arg_{\mathrm{last}}) f$ — функционал, вычисляющий первый (последний) ненулевой элемент горизонтальной строки.

В качестве второго признака возьмем ту же структуру, за исключение одного функционала $\mathrm{Hyper}\,P$, который заменим так:

$$\mathrm{Hyper}\,P(\omega, \varphi) = \arg_{\mathrm{last}} G(\omega, \varphi, r).$$

При этом заметим, что дополнительно производить сканирование для второго сенситивного признака не нужно: оба признака извлекаются из одной и той же гипертрейс 3TM и трейс TM матриц.

Далее определяются различные инвариантные признаки 3D модели, которые можно использовать для процедуры распознавания и классификации. В частности, ниже показан признак, который полностью инвариантен к группе движений и масштабированию, полученный по двум сенситивным к переносу признакам $\mathrm{SenTrans}_2(F)$ и $\mathrm{SenTrans}_1(F)$:

$$\mathrm{Res}(F) = \mathrm{SenTrans}_1(F) - \mathrm{SenTrans}_2(F).$$

Показанный выше признак, характеризует наибольшую ширину исходного объекта видимую под выбранным с помощью функционалов Θ, $\mathrm{Hyper}\,\Omega$ и $\mathrm{Hyper}\,\Theta$ углом обзора. Более строгое определение выглядит следующим образом: характеризует наибольшую длину между точками множества, полученными как ортогональное отображение всех вершин исходного 3D объекта на прямую, параллельную сетке сканирующих плоскостей и перпендикулярную выбранному направлению — вектору нормали к плоскости сечения.

Кроме полученного выше признака можно построить множество других признаков также инвариантных к группе движений и масштабированию.

Промежуточным продуктом вычислений является тройка $(\omega_i, \varphi_j, r_k)$, которой соответствует элемент матрицы с номером (i, j, k) и значением $\Pi(F_{\mathrm{sect}})$, который характеризует информативный признак фигуры, полученной в сечении объекта F плоскостью $B\left(\eta(\omega_i, \varphi_j), r_k\right)$. Поэтому параллельно в одной технике сканирования можно определить уравнения граничных плоскостей:

$$\cos\varphi_j \,\sin\omega_i \cdot x + \cos\varphi_j \,\cos\omega_i \cdot y + \sin\varphi_j \cdot z = r_k.$$

Стоит отметить, что, используя сортировку вектора по возрастанию и выбирая его любую i-ю координату, можно выбирать различные граничные плоскости под любым углом обзора, содержащие внутри себя исходный объект.

Таким образом, определив параметрически множество коэффициентов уравнений граней, можно полностью построить выпуклый многогранник, содержащий внутри себя исходный объект. Причем можно построить не только выпуклый параллелепипед, но и любой другой выпуклый многогранник. Чем больше признаков такого класса вычисляется, тем более полигональный становится данный многогранник, вплоть до точной копии самого объекта (аналогично, obj или 3ds модели).

Итак, одним из сильных сторон гипертрейс-преобразования является не только предобработка и анализ 3D изображения в одной технике сканирования, но и возможность его постобработки в той же технике. Другими словами, можно производить процесс, противоположный полигональному сглаживанию поверхности 3D объекта. При этом мы параллельно знаем не только размеры этого многогранника, но и его границы (параметрическая система уравнений плоскостей граней объекта).

Аналогично можно построить выпуклый многогранник, полностью содержащийся в исходной 3D модели.

Функционалы, инвариантные к повороту

Стандартный перебор всех углов сетки плоскостей ω и φ в топологическом смысле для непрерывного случая дает модель концентрических сфер с центром в начале координат. Каждой плоскости сопоставим точку ее касания с соответствующей сферой.

Рассмотрим единичную сферу. Множество точек на сфере образуют сетку, которую будем называть опорной. Отметим, что смена углов происходит согласно узлам опорной сетки, однозначно определяющим единственную плоскость.

В функциональном виде свойство инвариантности к повороту для любого функционала ζ выглядит следующим образом: $\zeta\left(f(M(\nu,\alpha)X)\right) = \zeta\left(f(X)\right)$ для всех допустимых $f(X)$, $\forall\nu$ и $\forall\alpha$, где X — точка объекта в трехмерном пространстве, $\nu(x, y, z)$ — единичный вектор (ось вращения), α — угол поворота вокруг вектора ν, $M(\nu, \alpha)$ — двумерная

матрица поворота:

$$M(\nu,\alpha)=\begin{pmatrix} \cos\alpha+x^2(1-\cos\alpha) & xy(1-\cos\alpha)-z\sin\alpha & xz(1-\cos\alpha)+y\sin\alpha \\ yx(1-\cos\alpha)+z\sin\alpha & \cos\alpha+y^2(1-\cos\alpha) & yz(1-\cos\alpha)-x\sin\alpha \\ zx(1-\cos\alpha)-y\sin\alpha & zy(1-\cos\alpha)+x\sin\alpha & \cos\alpha+z^2(1-\cos\alpha) \end{pmatrix}$$

Для дискретного случая на обычной карте глобуса вблизи полюса наблюдается более плотное скопление точек, чем у экватора. Поэтому если при повороте полюс совместить с точкой на экваторе, то будут заметны отклонения точек узлов исходной и повернутой сетки, что скажется на распознавании объектов, так как каждая точка опорной сетки на сфере однозначно определяет угол наклона сетки параллельных плоскостей. Поэтому необходимо строить такую опорную сетку, которая будет обладать равномерным распределением точек на сфере. Идеи построения равномерной сетки на сфере можно найти в [3].

Инвариантность к повороту получаемых признаков зависит не только от плотности совокупности сеток сканирующих плоскостей в пространстве, но и применения специального типа функционалов, которые обрабатывают горизонтальные строки и вертикальные столбцы матрицы 3TM.

Свойство инвариантности к сдвигу периодической функции f с периодом r для любого функционала ζ будет иметь вид: $\zeta\left(f(x+b)\right)=\zeta\left(f(x)\right)$ для всех допустимых $f(x)$ и $\forall b$.

Так как равномерная сетка на сфере неизоморфна равномерной сетке на плоскости, то при формировании 3D трейс-матрицы 3TM возникают трудности сохранения целостности структуры и порядка следования строк самой гипертрейс-матрицы. Так как порядок сечений (дискретная форма 3D объекта) не изменяется при повороте сетки сканирующих плоскостей, то глубинные строки сохраняют свой порядок следования элементов. Поэтому основная проблема состоит в сохранении порядка следования строк и столбцов друг за другом. Эта проблема легко решается, если определить правила формирования 3D трейс-матрицы и ключевые точки, которые сами могут выступать как сенситивные к повороту признаки. Данные опорные ключевые точки определяют начало отсчета, от которых начинается заполняться гипертрейс-матрица (например, по часовой стрелке в направлении отсчета от второй ключевой точки).

В общем случае в терминах трейс-матриц строки и столбцы матрицы 3TM сдвинутся на ω-ое и φ-ое число вперед или назад соответственно, в зависимости от знака углов и правила нумерации узлов опорной сетки. При этом порядок их следования друг за другом не изменится. Соответственно ее гипертрейс-образ (графическое представление гипертрейс-матрицы) будет сдвинут вдоль горизонтальной 0ω и вертикальной оси 0φ на соответствующее расстояние, равное углам поворота.

Стоит помнить, что наилучший способ нумерации узлов опорной сетки будет наиболее близкий к методу ее построения, когда построение узлов и их нумерация будет осуществляться в одной технике.

Более подробно свойства инвариантности к повороту для 3D трейс-преобразования можно найти в [6, 7], а для 2D трейс-преобразования в плоскости сечения — в [8].

Ниже приведен пример признака, инвариантного к операции ротации и трансляции и сенситивного к операции гомотетии:

$$\mathrm{Res}(F)=\mathrm{Hyper}\,\Theta\circ\mathrm{Hyper}\,\Omega\circ\mathrm{Hyper}\,P\circ\mathrm{Hyper}\,T(\Theta\circ P\circ T),$$

где $T(\theta,\rho) = \sum_t f(\theta,\rho,t)$; $P(\theta) = \left(\min\limits_i f(\theta,\rho_i) + \max\limits_i f(\theta,\rho_i)\right)/2$; $\Theta = c\sum_k f(\theta_k)/n$; $\operatorname{Hyper} T(\omega,\varphi,r) = G(\omega,\varphi,r)$; $\operatorname{Hyper} P(\omega,\varphi) = 1/\min\limits_r G(\omega,\varphi,r)$; $\operatorname{Hyper}\Omega(\varphi) = \max\limits_\omega G(\omega,\varphi)$; $\operatorname{Hyper}\Theta = \sum_\varphi G(\varphi)$.

Функционалы, сенситивные к повороту

Роль сеток при конструировании гипертриплетных признаков подробно описана в разделе выше. Использовать неравномерные опорные сетки на сфере, которые в общем случае дают признаки, сенситивные к повороту, не представляется возможным, так как уровень ошибки, получаемой отклонением узлов сетки друг от друга при повороте сферы, является неконтролируемой величиной и в общем случае может сильно искажать вычисляемые признаки.

Однако получить признаки, сенситивные к повороту, можно при помощи применения сенситивных функционалов. Данный класс признаков не только увеличивает разнообразие конструируемых информативных признаков, но и позволяет извлекать информацию о пространственной ориентации объекта.

Свойство инвариантности к сдвигу периодической функции f с периодом r для любого функционала ζ будет иметь вид: $\zeta\left(f(x+b)\right) = \zeta\left(f(x)\right) - k$ или $\zeta\left(f(x+b)\right) = \zeta\left(f(x)\right) + k$ для всех допустимых $f(x)$ и $\forall b$, где r — любое положительное действительное число, $k \equiv b \pmod r$ и $0 \leqslant k < r$.

Сенситивные к повороту функционалы можно неформально интерпретировать как операцию выбора точки на двумерной поверхности в трехмерном пространстве независимо от пространственной ориентации 3D объекта. Так, как указывалось в предыдущем разделе, ключевые точки могут играть роль сенситивных к повороту признаков. Ниже приведен пример одного из них, который вычисляет вектор нормали к одной из секущих плоскостей:

$$\operatorname{SenRot}_1(F) = (\operatorname{Hyper}\Theta; \operatorname{Hyper}\Omega; \operatorname{Hyper} P) \circ \operatorname{Hyper} T(\Theta \circ P \circ T).$$

Здесь $T(\theta,\rho) = \operatorname{median}\limits_t f(\theta,\rho,t)$; $P(\theta) = \sum_i f(\theta,\rho_i)/n$; $\Theta = \max\limits_k f(\theta_k)$; $\operatorname{Hyper} T(\omega,\varphi,r) = \Pi(F_{\text{sect}})$; $\operatorname{Hyper} P = \arg\max\limits_r G(\omega,\varphi,r)$; $\operatorname{Hyper}\Omega = \arg\max\limits_\omega G(\omega,\varphi,r)$; $\operatorname{Hyper}\Theta = \arg\max\limits_\varphi G(\omega,\varphi,r)$, где $median$ — медиана дискретного ряда чисел.

Ключевые точки отличаются друг от друга выбором функционала:

$$\operatorname{SenRot}_2(F) = \{\omega_2^*; \varphi_2^*; r_2^*\} = \left\{\arg\min_\omega G; \arg\min_\varphi G; \arg\min_r G\right\}.$$

Определив несколько таких векторов, мы получаем ключевые точки опорной сетки на сфере, предварительно нормализовав их длину до единицы. При этом стоит отметить, ускорить данную процедуру можно сделав значение r сразу равным единице. Длина между такими точками является признаком, инвариантным к группе движений и операции гомотетии исходного 3D объекта. Причем это может быть длина дуги, длина прямолинейного отрезка и т. п.:

$$\operatorname{Res}(F) = (\sin\varphi_2\cos\omega_2 - \sin\varphi_1\cos\omega_1)^2 + (\sin\varphi_2\sin\omega_2 - \sin\varphi_1\sin\omega_1)^2 + (\cos\varphi_2 - \cos\varphi_2)^2.$$

Угол между прямыми, проходящими через одинаково определяемую ключевую точку для повернутого и исходного объекта, даст значение угла, при котором один объект можно

совместить поворотом в другой. Данная процедура осуществляется в той же технике, что и сканирование и обработка 3D изображения и подробно в данной статье описываться не будет. Ее можно найти в [9]. Основная идея построения такого типа признаков заключается в выборе двух чисел из всего множества значения циклической функции, обладающих конкретными особенностями (например, максимум или медиана). После этого находится между ними расстояние, которое в общем случае постоянно вне зависимости от операций ротаций 3D объекта.

Функционалы, инвариантные к масштабированию

Стоит отметить, что одним из преимуществ пространственного трейс-преобразования является возможность конструирования признаков 3D фигур инвариантных одновременно не только к группе движений, но и к масштабированию. Свойство инвариантности к масштабированию для любого функционала ζ будет иметь вид: $\zeta\left(f(a\,x)\right) = \zeta\left(f(x)\right)$ для всех допустимых $f(x)$ и $\forall a$.

Получение данного класса признаков достигается несколькими способами. К первому способу можно отнести подбор специальных типов функционалов, которые дают одинаковое значение вне зависимости от масштаба объекта. Гипертриплетные признаки, построенные на их основе при применении также функционалов инваринтных к трансляции и ротации, дают инвариантное описание 3D объекта. Как правило, данные признаки $\mathrm{Res}(F)$ описывают свойства или особенности формы 3D объекта. В качестве примера можно привести такой функционал как число пересечений сканирующей плоскости с 3D изображением, который указывает на наличие пустых полостей внутри объекта. Данный функционал также способен определить количество впадин, пиков или составных частей объекта, число которых не изменяется при изменении масштаба. Другие виды функционалов могут определить углы наклона склонов холмов на поверхности тела, если таковые имеются, и т. п.

К таким типам функционалов можно отнести еще определение числа локальных максимумов (минимумов) функции, коэффициент корреляции и т.п.

Ко второму способу относятся, как правило, комбинации отношения признаков, которые нивелируют коэффициент масштабирования. Например, такие комбинации, как $\max(l)/\min(l)$, $\sqrt{S(F_{\mathrm{sect}})}/P(F_{\mathrm{sect}})$, $S^3(F)/V^2(F)$ и т. п., где l — длина отрезка прямой, заключенной в 2D фигуре сечения, $S(F_{\mathrm{sect}})$ — площадь сечения, $P(F_{\mathrm{sect}})$ — периметр сечения, $S(F)$ — площадь поверхности исходной 3D фигуры, $V(F)$ — объем сходной 3D фигуры. При конструировании такого класса гипертриплетных признаков необходимо учитывать размерность m вычисляемого признака. Так, длина отрезка имеет $m = 1$, а периметр сечения имеет $m = 2$. Так, чтобы снизить влияние линейных искажений и координатных шумов, следует выбирать и делить на такие признаки объектов, которые менее подвержены влиянию шума, чем другие признаки. Например, для признаков разных размерностей это будут радиус описанной сферы около объекта, площадь поверхности объекта и объем исходного тела соответственно.

Учитывая возможность использования различных видов функционалов, саму композиционную структуру гипертриплетного признака (перестановки функционалов местами), а также различные виды комбинаций отношений между признаками, можно получить большое количество признаков, инвариантных к масштабированию. Ниже представлен признак, инвариантный к операциям трансляции, ротации и гомотетии:

$$\mathrm{Res}(F) = \mathrm{Hyper}\,\Theta \circ \mathrm{Hyper}\,\Omega \circ \mathrm{Hyper}\,P \circ \mathrm{Hyper}\,T(\Theta \circ P \circ T),$$

где $T(\theta,\rho) = \min\limits_{t} f(\theta,\rho,t)$; $P(\theta) = \sqrt{\sum_j (f(\theta,\rho_j) - \sum_i f(\theta,\rho_i))}/n$; $\Theta = \sum_i f(\theta_i)$; $\operatorname{Hyper} T = \Pi(F_{\text{sect}})$; $\operatorname{Hyper} P(\omega,\varphi) = \sum_r G(\omega,\varphi,r)$, если $G \geqslant \operatorname{median}\limits_{r}(G)$; $\operatorname{Hyper}\Omega(\varphi) = \max\limits_{\omega}(G(\omega,\varphi))$; $\operatorname{Hyper}\Theta = \operatorname{kol}\limits_{\varphi}(G(\varphi))$, если $G \leqslant \sum_\varphi G(\varphi)/n$.

Функционалы, сенситивные к масштабированию

Свойство сенситивности к масштабированию для любого функционала ζ будет иметь вид: $\zeta(f(a\,x)) = (1/a)\zeta(f(x))$ или $\zeta(f(a\,x)) = a\zeta(f(x))$ для всех допустимых $f(x)$ и $\forall a \neq 0$.

Сенситивность получаемых признаков $\operatorname{Res}(F)$ достигается путем подбора функционалов, чувствительных к масштабированию, удовлетворяющих условию выше.

На языке трейс-матриц сенситивность выражается в том, что гипертрейс-матрица увеличивает $\mu > 1$ количество своих ненулевых значений в i-ой строке глубины, т. к. исходный объект будет пересекать больше секущих плоскостей. Тем самым соответствующие графические образы трейс-матриц расширяются при $\mu > 1$ относительно оси, определяющей расстояние до фигуры. Для трейс-матриц 2D изображений в плоскости сечения эту роль играет ось ρ, для трейс-матриц 3D изображений — ось r.

Учет свойства сенситивности может быть произведен двумя способами. В первом случае все получаемые признаки приводятся к одному масштабу (например, единичному). Так, чтобы получить единичный масштаб, достаточно разделить все вычисляемые признаки, например, на максимальный элемент. Для получения более устойчивого значения признака к помехам необходимо брать несколько определяющих элементов по разным правилам. Далее, зная и учитывая коэффициент масштабирования, вычисленные гипертриплетные признаки исходного объекта сравниваются с признаками других объектов:

$$\sum_{j,j\neq k} \left| \frac{\operatorname{Res}(F)_{i,k}}{\mu_{i,k}^{m[i,k]}} - \frac{\operatorname{Res}(F)_{i,j}}{\mu_{i,j}^{m[i,j]}} \right| \leqslant \delta_i,$$

где j — номер объекта; $m[i,j]$ — размерность i го признака j-го объекта; $\mu_{i,j}$ — коэффициент масштабирования i го признака j-го объекта; δ_i — порог схожести объектов по i му признаку.

Коэффициент масштабирования может быть получен непосредственно из результатов техники сканирования. Так, среди всех сечений $F_{\text{sect}\,i,j}$ под j-м углом i-й плоскости выбирается определяющее сечение $Q_k = h(S_k)_{i,j}$ по правилу h отдельно для каждого признака, где S — вектор N признаков сечения $F_{\text{sect}\,i}$. В качестве правила h может быть взята функция среднего арифметического для получения, например, сечения с усредненным значением периметра. Затем все значения k-го признака делятся на это значение: $\mu_k = (F_{sectk})_{i,j}/Q_k$.

Второй подход учета сенситивных признаков состоит в сведении всех получаемых признаков к неопределенному масштабу — безразмерной величине. Ниже приведен пример признака, сенситивного к масштабированию и инвариантного к группе движений:

$$\operatorname{SenGomo}(F) = \operatorname{Hyper}\Theta \circ \operatorname{Hyper}\Omega \circ \operatorname{Hyper} P \circ \operatorname{Hyper} T(\Theta \circ P \circ T),$$

где $T(\theta,\rho) = \max\limits_{t} f(\theta,\rho,t)$; $P(\theta) = \Delta t \sum_i f(\theta,\rho_i)$; $\Theta = \sum_i f(\theta_i)$; $\operatorname{Hyper} T = \Pi(F_{\text{sect}})$; $\operatorname{Hyper} P(\omega,\varphi) = \operatorname{median}\limits_{r} G(\omega,\varphi,r)\Delta h$; $\operatorname{Hyper}\Omega(\varphi) = \max\limits_{\varphi} G(\omega,\varphi)$; $\operatorname{Hyper}\Theta = \min\limits_{\varphi} G(\varphi)$; Δt — расстояние между параллельными прямыми в плоскости сечения, Δh — расстояние между параллельными плоскостями.

Результаты. Ввиду того, что статья носит концептуальный характер, описывает математическую модель и интеллектуальные возможности метода, а также того, что объем настоящей работы ограничен, реальные практические эксперименты и тестирование признаков на различных базах 3D объектов в данной статье не проводится. Их можно найти в [10, 11].

По мере необходимости в данной статье приводилось не только множество теоретических примеров признаков, обладающие указанными свойствами, но и разбирались и описывались дополнительные возможности метода, улучшающие интеллектуальную обработку и анализ 3D изображений.

В завершении подчеркнем, что для получения признаков, инвариантных к группе движений и масштабированию, необязательно использовать только такие функционалы, которые инвариантны к указанным выше операциям. Инвариантность может быть достигнута даже при использовании части функционалов, сенситивных к некоторым преобразованиям 3D изображений.

Заключение

Метод, рассмотренный в статье, обладает определенной универсальностью, так как схема сканирования не привязана к геометрическим особенностям исходной модели. В связи с этим предлагаемая методика ориентирована на объекты любой сложности и конфигурации, что повышает интеллектуальные возможности 3D трейс-преобразования.

Предлагаемый метод позволяет давать инвариантное описание исходного 3D объекта при любых операций трансляции, ротации и гомотетии. Помимо того, конструируемые гипертриплетные признаки обладают высокой устойчивостью к координатному шуму и линейным искажениям [12], они имеют хорошую геометрическую интерпретацию, улучшающие интеллектуальные возможности анализа. Высококачественный интеллектуальный анализ с помощью гипертрейс-преобразования реализуется благодаря возможности регулирования свойств построенных признаков. Сенситивные признаки, кроме того, дают возможность извлекать в той же технике сканирования множество дополнительной информации об объекте, в частности его преобразованиях, знание которой необходимо при решении некоторого класса практических задач (например, позиционирование инструмента в робототехнике").

Также одной из интеллектуальных способностей предлагаемого метода является высокоуровневая предобработка, обработка и анализ 3D изображения в одной технике сканирования.

Авторы планируют развить данный метод для на анализа не только бинарных и монохромных 3D изображений, но и цветных и текстурных 3D изображений. Аналогичные результаты уже были получены при анализе цветных и текстурных 2D изображений в [13, 14]. Интеллектуальный уровень гипертрейс-преобразования может быть повышен благодаря развитию теории на аффинно-инвариантные преобразования. Аналогичный результат для трейс-преобразования уже получен в [15]. Планируеся также развить гипертрейс-преобразования для интеллектуального анализа и распознавания деформированных и поврежденных 3D объектов. Сейчас авторами ведется работа по развитию теории при нелинейной фильтрации 3D изображений на этапе его предобработки в той же технике сканирования, аналогично результатам для 2D изображений [16].

Литература

[1] *Федотов Н. Г., Семов А. А.* Краткий обзор основных подходов к анализу 3D-моделей и разработка 3D трейс-преобразования // *Новые информационные технологии и системы (НИТИС-2012): Сб. тр. X Междунар. научно-технич. конф.* Пенза: Изд-во Пенз. ГУ, 2012. С. 222–225.

[2] *Федотов Н. Г., Фионов Н. С., Романенко Ю. А., Баннов В. Я.* Развитие принципов интеллектуального поиска биометрических изображений на основе стохастической геометрии и функционального анализа // *Надежность и качество: Тр. Междунар. симпозиума* / Под ред. Н. К. Юркова. Пенза: Изд-во Пенз. ГУ, 2012. Т. 2. С. 394.

[3] *Федотов Н. Г., Семов А. А.* Идеи построения равномерной сетки на сфере и 3D трейс-преобразование // *Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XIII Междунар. научно-технич. конф.* Пенза: Изд-во АННОО «Приволжский Дом знаний», 2013. С. 23–26.

[4] *Федотов Н. Г., Семов А. А.* 3d трейс-преобразование и его свойства // *XXI век: итоги прошлого и проблемы настоящего — плюс. Научно-методический журнал. Серия: технические науки. Информационные технологии.* Пенза: Изд-во Пенз. гос. технол. Акад., 2013. № 10(14). С. 68–74.

[5] *Fedotov N. G.* The Theory of image — recognition features based on stochastic geometry // *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications*, 1998. Vol. 8, No. 2. P. 264–266.

[6] *Семов А. А.* Об одном подходе к распознаванию 3D-изображений // *Надежность и качество: Тр. Междунар. симпозиума* / Под ред. Н. К. Юркова. Пенза: Изд-во Пенз. ГУ, 2013. Т. 1. С. 350–351.

[7] *Fedotov N. G., Ryndina S. V., Syemov A. A.* Trace transform of spatial images // *11th Conference (International) on Pattern Recognition and Image Analasis: New Information Technologies (PRIA-11-2013) Proceedings. (V. I-II).* Samara: IPSI RAS, 2013. Vol. I. P. 186–189.

[8] *Федотов Н. Г., Семов А. А.* Гипертрейс-преобразование, инвариантное к группе движений 3D-объектов // *Современные методы и средства обработки пространственно-временных сигналов: Сб. статей XII Всеросс. научно-технич. конф.* Пенза: Изд-во АННОО «Приволжский Дом знаний», 2014. С. 38–43.

[9] *Федотов Н. Г.* Анализ свойств триплетных признаков распознавания при различных вариантах сканирования изображений // *Надежность и качество: Тр. Междунар. симпозиума* / Под ред. Н. К. Юркова. Пенза: Изд-во Пенз. ГУ, 2013. Т. 1. С. 80–82.

[10] *Федотов Н. Г., Семов А. А.* Новый метод распознавания и поиска 3D-объектов по базам данных // *Открытые инновации — вклад молодежи в развитие региона: Сб. материалов регионального молодежного форума.* Пенза: Изд-во Пенз. ГУ, 2013. Т. 1. С. 192–194.

[11] *Федотов Н. Г., Семов А. А., Крючкова Е. А.* Особенности реализации и способы ускорения вычислений 3D-трейс-преобразования // *Надежность и качество: Тр. Междунар. симпозиума* / Под ред. Н. К. Юркова. Пенза: Изд-во Пенз. ГУ, 2014. Т. 1. С. 393–396.

[12] *Семов А. А.* Повышение надежности распознавания 3D-объектов на основе методов стохастической геометрии // *Надежность и качество: Тр. Междунар. симпозиума* / Под ред. Н. К. Юркова. Пенза: Изд-во Пенз. ГУ, 2014. Т. 1. С. 393–396.

[13] *Fedotov N. G., Mokshanina D. A.* Recognition of halftone textures from the standpoint of stochastic geometry and functional analysis // *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications*, 2010. Vol. 20, No. 4. P. 551–556.

[14] *Fedotov N. G., Mokshanina D. A.* Recognition of images with complex half-tone texture // *Measurement Techniques*, 2011. Vol. 53, No. 11. P. 1226–1232.

[15] *Федотов Н. Г.* Теория признаков распознавания образов на основе стохастической геометрии и функционального анализа. М.: ФИЗМАТЛИТ, 2009. 304 с.

[16] *Федотов Н. Г., Крючкова Е. А., Моисеев А. В., Семов А. А.* Предварительная обработка изображений на основе трейс-преобразований // *Надежность и качество: Тр. Междунар. симпозиума* / Под ред. Н. К. Юркова. Пенза: Изд-во Пенз. ГУ, 2011. Т. 2. С. 315–316.

References

[1] *Fedotov N. G., Syemov A. A.* 2012. Brief review of approaches for analysis of 3D-models and development of 3D trace-transform. *New information technologies and systems (NITIS-2012): Proceedings of the X International Scientific and Technical Conference*. Penza: Publ. PenzGU. 222–225. (in Russ.)

[2] *Fedotov N. G., Fionov N. S., Romanenko Yu. A., Bannov V. J.* 2012. Developing the principles for intelligent search of biometric images based on stochastic geometry and functional analysis. *Reliability and quality: Proceedings of the International Symposium, ed. N. K. Jurkov*. Penza: Publ. PenzGU. 2:394. (in Russ.)

[3] *Fedotov N. G., Syemov A. A.* 2013. The ideas of building a uniform grid on the sphere and 3d Trace transform. *Problems of Informatics in Education, Management, Economics and Technology: Proceedings of the XIII International Scientific and Technical Conference*. Penza: Publ. ANPSEO "Privolzskiy Dom Znaniy". 23–26. (in Russ.)

[4] *Fedotov N. G., Syemov A. A.* 2013. The 3d trace-transform and its properties. *XXI century: past results and present problems – plus. Scientific-methodical journal. Series: engineering science. Information technology*. Penza: Publ. Penz. Gov. Technol. Academ. 10(14): 68–74. (in Russ.)

[5] *Fedotov N. G.* 1998. The Theory of Image-Recognition Features Based on Stochastic Geometry. *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications* 8(2):264–266.

[6] *Syemov A. A.* 2013. About an approach to the recognition of 3D images. *Reliability and quality: Proceedings of the International Symposium, ed. N. K. Jurkov*. Penza: Publ. PenzGU. 1:350–351. (in Russ.)

[7] *Fedotov N. G., Ryndina S. V., Syemov A. A.* 2013. Trace transform of spatial images. *11-th International conference on Pattern Recognition and Image Analasis: New Information technologies (PRIA-11-2013). Conference Proceedings (V. I-II)*. Samara: IPSI RAS. 1:186–189.

[8] *Fedotov N. G., Syemov A. A.* 2014. Hypertrace-transform invariant to the motions group of 3D-objects. *Modern methods and tools the processing of spatio-temporal signals: Proceedings of the XII All-Russian Scientific and Technical Conference*. Penza: Publ. ANPSEO "Privolzskiy Dom Znaniy". 38–43.

[9] *Fedotov N. G.* 2013. Analyzing the properties of the triplet recognition features in different types of images scanning. *Reliability and quality: Proceedings of the International Symposium, ed. N. K. Jurkov*. Penza: Publ. PenzGU. 1:80–82. (in Russ.)

[10] *Fedotov N. G., Syemov A. A.* 2013. A new method of 3D-objects recognition and retrieval from databases. *Open innovations - the contribution of the youth in the region development: Proceedings of regional youth forum (V. I-II)*. Penza: Publ. PenzGU. 1:192–194. (in Russ.)

[11] *Fedotov N. G., Syemov A. A., Kryuchkova E. A.* 2014. Some particularities of 3D trace-transform implementation, and the ways to accelerate their calculations. *Reliability and quality: Proceedings of the International Symposium, ed. N. K. Jurkov*. Penza: Publ. PenzGU. 1:390–393. (in Russ.)

[12] *Syemov A. A.* 2014. Reliabilization of 3D objects' recognition on the basis of stochastic geometry methods. *Reliability and quality: Proceedings of the International Symposium, ed. N. K. Jurkov*. Penza: Publ. PenzGU. 1:393–396. (in Russ.)

[13] *Fedotov N. G., Mokshanina D. A.* 2010. Recognition of halftone textures from the standpoint of stochastic geometry and functional analysis. *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications* 20(4):551–556.

[14] *Fedotov N. G., Mokshanina D. A.* 2011. Recognition of images with complex half-tone texture. *Measurement Techniques*. 53(11):1226–1232.

[15] *Fedotov N. G.* 2009. The theory of patterns recognition features based on stochastic geometry and functional analysis. Moscow: PhysMathLit. 304 p. (in Russ.)

[16] *Fedotov N. G., Kryuchkova E. A., Moiseev A. V., Syemov A. A.* 2011. Images preprocessing based on trace-transform. *Reliability and quality: Proceedings of the International Symposium, ed. N. K. Jurkov*. Penza: Publ. PenzGU. 2:315–316. (in Russ.)

О некоторых вопросах анализа пучков временных рядов*

***Н. В. Филипенков*[1], *М. А. Петрова*[2]**
[1]n.filipenkov@mail.ru, [2]marina_petrova@mail.ru
[1]САС институт, Москва, ул. Станиславского, 21-1; [2]НИЯУ МИФИ, Москва, Каширское ш., 31

В настоящей работе рассматривается разрабатываемый авторами подход к поиску закономерностей в пучках нестационарных k-значных временных рядов. Этот подход позволяет выявлять закономерности, которые подвергаются «плавным» структурным изменениям с течением времени.

Настоящая работа посвящена описанию результатов апробации разрабатываемого подхода на модельных и реальных задачах. Испытания на модельных задачах показали, что подход позволяет эффективно находить заложенные закономерности при достаточно высоком уровне шума. Эксперименты на модельных пучках временных рядов показали, что использование меры сходства закономерностей в функционале качества существенно повышает точность прогнозирования. В рамках экспериментов был получен диапазон весов, при котором достигается максимальное качество распознавания. Анализ реальных временных рядов с применением разрабатываемого алгоритма свидетельствовал об эффективности алгоритма при краткосрочном прогнозировании. Вместе с тем алгоритм решает и задачу интеллектуального анализа данных, предлагая закономерности, описывающие взаимосвязь одномерных временных рядов.

Таким образом, апробация разрабатываемого подхода к прогнозированию процессов с плавно меняющимися закономерностями на модельных и реальных данных позволяет судить о достаточной эффективности разрабатываемых авторами алгоритмов при анализе пучков временных рядов с плавно меняющимися закономерностями.

Ключевые слова: временные ряды; интеллектуальный анализ данных; мера сходства закономерностей; вычислительный алгоритм

On the analysis of multidimensional time series*

***N. V. Filipenkov*[1], *M. A. Petrova*[2]**
[1]SAS Institute, 21-1 Stanislavskogo Str., Moscow; [2]MEPhI, 31 Kashirskoye Sh., Moscow

In this paper, an approach for discovering rules in nonstationary finite-valued multidimensional time series is discussed. It allows one to discover rules that slightly change their structure over time. A measure of rule similarity is introduced and studied as a weight on the graph of rules.

This paper focuses on the results of the application of the discussed algorithm to the modeled and real problems. The experiments on the model problems show that the approach allows to mine the hidden rules efficiently even under high noise conditions. The experiments on the modeled multidimensional time series show that using the rules similarity measure in the quality function significantly increases the forecast accuracy. During the experiments, the

*Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00293.

weight range for maximum data mining quality was identified. The analysis of real time series based on the discussed approach show the algorithm's efficiency for short-term forecasting. In addition to that, the algorithm solves the data mining problem while finding the rules describing the interconnection of the univariate time series.

The application of the discussed approach for forecasting the processes with slightly changing rules on modeled and real data shows the efficiency of the developed algorithms for the analysis of multidimensional time series with slightly changing rules.

Keywords: time series; data mining; similarity measure; algorithm

Введение

В настоящее время анализ временных рядов является крайне актуальной задачей в различных сферах деятельности человека: медицине, экономике, физике, кибернетике. При этом часто возникает необходимость исследования сразу нескольких процессов или показателей одного процесса в их взаимосвязи и взаимовлиянии — изучения пучков временных рядов. Пучки временных рядов могут, например, описывать процессы жизнедеятельности человека, стоимость акций на бирже, курсы валют и т. д. Пучок временных рядов, учитывая множество характеристик явления, позволяет описать процесс или систему процессов наиболее полно, что, в свою очередь, позволяет сделать более точный прогноз. Возможность системного анализа процессов, их более точного описания определила высокий интерес исследователей к изучению пучков временных рядов [1 — 12].

Изучение пучков временных рядов подразумевает не только прогнозирование значений рядов, но предполагает решение задачи в рамках области интеллектуального анализа данных. Это означает, что необходимо выявить и описать закономерности, определяющие поведение временных рядов. Найденные закономерности могут быть представлены в виде уравнений [2, 4, 5], ассоциативных [6, 7], «эпизодических» [8] или прочих правил [10, 11, 9].

В большинстве реальных задач измерения проводятся в дискретные моменты времени, поэтому во многих работах рассматриваются дискретные временные ряды. При этом в работах, посвященных поиску правил [6, 10, 11, 9], рассматриваются пучки, где значениями временных рядов являются элементы некоторого конечного алфавита. Для поиска правил в «непрерывных» временных рядах используются методы дискретизации или символьного представления [10, 12].

Пучок временных рядов отражает характеристики явления во времени, но само явление может меняться с течением времени. Нестационарными в общем смысле называются временные ряды, свойства которых непостоянны во времени. Во многих областях такие ряды составляют большинство, так как почти все явления под воздействием различных факторов претерпевают изменения. Для анализа нестационарных временных рядов был предложен целый ряд адаптивных методов: экспоненциальное сглаживание и его модификации [13, 14], модели семейства ARIMA [2], модели семейства ARCH [4, 15], множественная регрессия [5], модели, основанные на использовании спектральных характеристик рядов [16, 17].

Основные определения *Пучком временных рядов* $\mathfrak{S}$ называется совокупность взаимосвязанных временных рядов S_i, $i \in \{1, 2 \ldots, N\}$. Каждый ряд S_i представляет собой последовательность значений конечнозначной логики E_{k_i}. Каждому элементу ряда соответствует некоторый момент времени, и эти моменты времени для всех рядов одинаковы. Поэтому одинакова и длина всех рядов, которая обозначается через T. Таким образом,

пучок временных рядов $\mathfrak{S}$ есть матрица размера $N \times T$, где элемент i-й строки принадлежит множеству E_{k_i}. Значения ряда $S_i, i \in \{1, 2 \ldots, N\}$ в момент времени $t \in \{1, 2 \ldots, T\}$ обозначим через $a(i,t)$ или $a_{i,t}$.

Маской ω на прямоугольнике $N \times \Delta$ назовем булеву матрицу размера $N \times \Delta$ (здесь параметр Δ определяет максимальный отступ по времени). Число единиц в маске ω будем называть *мощностью* маски и обозначать через $||\omega||$. Элемент маски, находящийся в i-й строке и j-м столбце, будем обозначать через $\omega(i,j)$ или $\omega_{i,j}$. *Закономерностью* R назовем набор (p, ω, f) с такими особенностями:

1) число $p \in \{1, 2, \ldots, N\}$ указывает на целевой ряд (ряд, значения которого определяются закономерностью R);
2) маска ω указывает на значения рядов, являющиеся аргументами функции f;
3) частично определенная функция f задает зависимость значений целевого ряда от переменных, на которые указывает маска ω.

$$f\colon E_{k_{i_1}} \times \ldots \times E_{k_{i_{||\omega||}}} \to E_{k_p} \cup \{\lambda\},$$

где $\omega(i_1, j_1), \ldots, \omega(i_{||\omega||}, j_{||\omega||})$ — единичные элементы матрицы ω; p — номер целевого ряда; символ λ обозначает, что f не определена на соответствующем наборе значений переменных.

Если значения всех рядов представляют собой числа k-значной логики ($E_{k_1} = \cdots = E_{k_N} = E_k$), то функция f принадлежит множеству P_k^* всех частично определенных функций k-значной логики.

Задача состоит в поиске закономерностей и прогнозировании. Найденные закономерности позволяют прогнозировать значения целевого ряда, делать выводы о характере зависимостей между рядами, моделировать целевой ряд или весь пучок временных рядов.

В предыдущих работах авторов [18, 19] рассматриваются алгоритм поиска постоянных закономерностей, по аналогии с [6, 7, 20] вводятся понятия достоверности $\mathrm{Conf}\,(R, \mathfrak{S})$ и поддержки $\mathrm{Supp}\,(R, \mathfrak{S})$ закономерности R на пучке $\mathfrak{S}$, оценивается необходимая длина пучка временных рядов, вводится понятие системы закономерностей, исследуется понятие её полноты, объясняется, каким образом подход позволяет применять к построенным закономерностям конструкции алгебраического подхода, предложенные в работах Ю. И. Журавлева [21] и К. В. Рудакова [22]. В работах производится построение меры сходства закономерностей, определяется понятие изменяющейся закономерности.

Изменяющейся закономерностью $\tilde{R}$ для последовательности отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ на пучке временных рядов $\mathfrak{S}$ называется система закономерностей $R^1, \ldots, R^m$, где каждая закономерность взаимно однозначно соответствует некоторому отрезку $\mathfrak{S}^i$, $i = 1, 2, \ldots, m$. Вообще говоря, отрезки могут пересекаться между собой. Будем называть стационарные закономерности $R^1, \ldots, R^m$ *шагами*, которые *составляют* изменяющуюся закономерность $\tilde{R}$.

В работах [18, 19] описывается поиск плавно меняющихся закономерностей на графе (рис. 1).

Вершинами графа являются стационарные закономерности, найденные на каждом из отрезков, а также две дополнительные вершины: beg и end. С каждой вершиной ассоциированы показатели качества закономерности. Дугами на графе связаны закономерности соседних отрезков, что отражает факт возможного «превращения» одной закономерности в другую. С каждой дугой ассоциирован вес — мера сходства соответствующих закономерностей. Веса дуг, соединяющие закономерности крайних отрезков с вершинами *beg* и *end*, полагаются равными нулю.

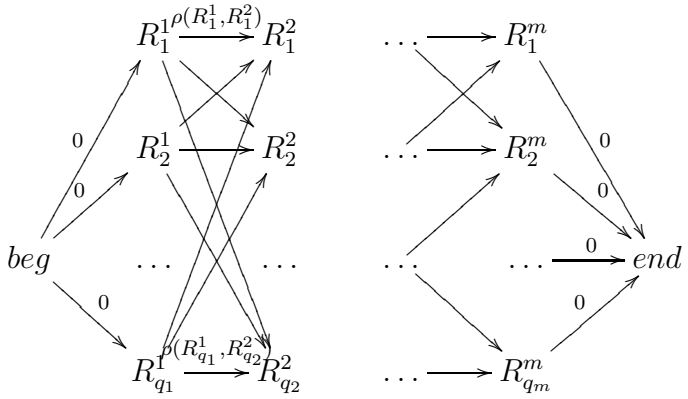


Рис. 1. Граф закономерностей

Задача выделения наилучшей изменяющейся закономерности состоит в поиске пути между вершинами beg и end на ориентированном графе, который максимизирует показатели качества закономерностей вершин, входящих в него, и минимизирует суммарный вес ребер.

Эта задача сводится к стандартной задаче поиска кратчайшего пути на графе, если использовать в качестве веса вершины величину $(1-Q_{\text{step}})$, где Q_{step} — *функционал качества шага изменяющейся закономерности* $\tilde{R}$, который задается следующим образом:

$$
\begin{aligned}
&Q_{\text{step}}(R_i^j, R_l^{j+1}) = w_{\text{conf}}\text{Conf}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right) + w_{\text{supp}}\text{Supp}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right) + w_{\text{similarity}}(1-\rho(R_i^j, R_l^{j+1}));\\
&Q_{\text{step}}(\text{beg}, R_i^j) = 0;\\
&Q_{\text{step}}(R_i^j, \text{end}) = w_{\text{conf}}\text{Conf}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right) + w_{\text{supp}}\text{Supp}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right),\\
&\qquad\qquad j = 1, 2, \ldots, m-1;\ \ i = 1, 2, \ldots, q_j;\ \ l = 1, 2, \ldots, q_{j+1}.
\end{aligned}
$$

Здесь $\text{Conf}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right)$ и $\text{Supp}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right)$ — показатели качества закономерности; $\rho(R_i^j, R_l^{j+1})$ — мера сходства закономерностей, детально описанная в работе [19]. Веса w_{conf}, w_{supp} и $w_{\text{similarity}}$ функционала качества шага удовлетворяют следующим условиям:

$$
\begin{aligned}
0 \leqslant w_{\text{conf}} &\leqslant 1;\\
0 \leqslant w_{\text{supp}} &\leqslant 1;\\
0 \leqslant w_{\text{similarity}} &\leqslant 1;\\
w_{\text{conf}} + w_{\text{supp}} + w_{\text{similarity}} &= 1.
\end{aligned}
$$

В работе [19] доказывается, что для произвольных закономерностей R_1, R_2 верно неравенство:

$$0 \leqslant Q_{\text{step}}(R_1, R_2) \leqslant 1.$$

Таким образом, вес вершины $(1 - Q_{\text{step}})$ является неотрицательным и для решения задачи поиска кратчайшего пути на графе удобно использовать стандартные алгоритмы:

1) алгоритм Дейкстры [23] со сложностью $\underline{\underline{O}}(n^2)$ или его реализацию с фибоначчиевой кучей со сложностью $\underline{\underline{O}}(n\log n)$, где n — число вершин графа;
2) алгоритм поиска кратчайшего расстояния в топологически отсортированном графе [23] со сложностью $\underline{\underline{O}}(n^2)$, где n — число вершин графа.

Изменяющуюся закономерность $\tilde{R}$ будем называть *плавно меняющейся*, если она составлена из закономерностей, лежащих на кратчайшем пути из вершины beg в вершину end и выполнено неравенство $w_{similarity} > 0$ для веса меры сходства закономерностей функционала качества шага.

Примеры решения модельных задач. С целью испытания предложенного подхода для решения практических задач был подготовлен экспериментальный стенд. Стенд позволяет импортировать и генерировать временные ряды, проводить поиск стационарных и изменяющихся закономерностей, а также решать задачи прогнозирования. С использованием стенда было проведено несколько серий экспериментов. Обозначения для параметров экспериментов представлены в табл. 1 и 2.

Таблица 1. Обозначения параметров

Обозначение	**Параметр**
Параметры генерации	
K	Значность пучка временных рядов
N	Количество рядов в пучке
T	Длина рядов
Δ_{gen}	Максимальный отступ по времени
$\|\|\omega_1\|\|$	Мощность маски первой закономерности
p_{gen}	Индекс целевого ряда
m_{gen}	Количество сегментов
ξ_{mask}	Количество изменений маски при переходе к следующему отрезку
π_{mask}	Вероятность каждого изменения маски при переходе к следующему отрезку
ξ_{func}	Доля изменяемых значений функции при переходе к новому отрезку
π_{func}	Вероятность каждого изменения функции при переходе к следующему отрезку
ε	Уровень шума (доля значений целевого ряда, определяемых случайно)
Параметры поиска стационарных закономерностей	
p_{mine}	Индекс целевого ряда
Δ_{mine}	Максимальный отступ по времени
μ	Максимальный вес маски
$\min \text{supp}_{\text{set}}$	Минимальная поддержка набора
Valid	Доля отрезка, которая используется для валидации закономерностей

Проводились две серии экспериментов на модельных рядах с целью выявить условия для наиболее эффективного применения предложенных в [19] алгоритмов интеллектуального анализа временных рядов.

Таблица 2. Обозначения параметров

Обозначение	Параметр
Фильтры базы знаний	
$\mathrm{conf}_{\min}$	Минимальная достоверность на обучении закономерности для включения в базу знаний
$err^{\mathrm{valid}}_{\max}$	Максимальная ошибка на валидации
$\mathrm{supp}_{\min}$	Минимальная поддержка закономерности для включения в базу знаний
Параметры поиска меняющихся закономерностей	
m_{mine}	Количество сегментов
v	Стоимость перемещения аргумента по вертикали (используется в мере сходства масок)
h	Стоимость перемещения аргумента по горизонтали (используется в мере сходства масок)
w_λ	Расстояние до значения λ (используется в мере сходства функций)
$\varkappa_{\mathrm{mask}}$	Вес меры сходства масок (используется в мере сходства закономерностей)
$\varkappa_{\mathrm{func}}$	Вес меры сходства функций (используется в мере сходства закономерностей)
w_{conf}	Вес меры, характеризующей точность закономерности (используется в функционале качества закономерностей)
w_{supp}	Вес достоверности (используется в функционале качества закономерностей)
$w_{\mathrm{similarity}}$	Вес меры сходства закономерностей (используется в функционале качества закономерностей)

В первой серии модельных экспериментов было проведено исследование влияния уровня шума в моделируемых пучках временных рядов на качество распознавания. Для каждого значения уровня шума проводилась серия из 100 экспериментов.

В каждом эксперименте в соответствии с параметрами K (значность), N (количество рядов) и T (длина рядов) генерировался пучок временных рядов. Все ряды, за исключением целевого, генерировались случайным образом при равномерном распределении. Пучок разбивался по времени на m_{gen} равных сегментов.

Затем генерировалась плавно меняющаяся закономерность, состоящая из m_{gen} шагов — стационарных закономерностей. Стационарная закономерность первого отрезка создавалась случайным образом в соответствии с условиями на ширину маски (Δ_{gen}) и количество аргументов ($||\omega_1||$). Каждый следующий шаг плавно меняющейся закономерности был получен из предыдущего путем сдвига ξ_{mask} элементов маски, где каждый сдвиг происходил с вероятностью π_{mask}. Вместе с тем при генерации следующего шага плавно меняющейся закономерности изменялась часть значений функции, определяемая долей ξ_{func} от общего количества наборов, на которых определена функция. Вероятность каждого изменения задавалась долей π_{func}. Значения параметров генерации представлены в табл. 3. Таким

образом, за счет плавных изменений стационарных закономерностей получалась плавно меняющаяся закономерность.

Целевой ряд заполнялся с использованием сгенерированной плавно меняющейся закономерности на основе значений других рядов. При этом каждое значение целевого ряда с вероятностью ε генерировалось случайным образом, а не в соответствии с закономерностью. Таким образом, в целевом временном ряде учитывался заданный уровень шума ε.

После этапа генерации пучка временных рядов в каждом эксперименте производился поиск плавно меняющихся закономерностей в сгенерированном пучке временных рядов. Значения параметров алгоритмов поиска закономерностей представлены в табл. 3 и 4. В соответствии с параметрами функционала качества шага изменяющейся закономерности (w_{conf}, w_{supp} и $w_{\text{similarity}}$) определялась наилучшая изменяющаяся закономерность, которая затем сравнивалась со сгенерированной закономерностью.

Таблица 3. Значения параметров в экспериментах

Параметр	**Серия 1**	**Серия 2**
Параметры генерации		
K	4	4
N	10	10
T	1000	1000
Δ_{gen}	20	20
$\|\|\omega_1\|\|$	3	2
p_{gen}	0	0
m_{gen}	3	3
ξ_{mask}	1	1
π_{mask}	1	1
ξ_{func}	0,03	0,03
π_{func}	1	1
ε	изменяется	изменяется
Параметры поиска стационарных закономерностей		
p_{mine}	0	0
Δ_{mine}	20	20
μ	5	4
$\min \text{supp}_{\text{set}}$	0	0
Valid	20%	20%

Критерии успешного эксперимента были определены следующим образом. Генерируемая стационарная закономерность и найденная стационарная закономерность называются *совпадающими*, если полностью совпадают их маски и доля различных значений функции не превышает 5% от общего числа наборов, на которых определены функции. *Изменяющиеся* закономерности называются *совпадающими*, если совпадают все их соответствующие шаги — стационарные закономерности. Эксперимент признается *успешным*, если найденная изменяющаяся закономерность оказывается совпадающей с генерируемой.

Для каждого значения уровня шума рассчитывалась доля успешных экспериментов по отношению к общему числу экспериментов при данном уровне шума.

Таблица 4. Значения параметров в экспериментах

Параметр	Серия 1	Серия 2
Фильтры базы знаний		
$\text{conf}_{\min}$	0,3	0,3
$\text{err}^{\text{valid}}_{\max}$	1,0	1,0
$\text{supp}_{\min}$	0	0
Параметры поиска меняющихся закономерностей		
m_{mine}	3	3
v	1	1
h	1	1
w_λ	4	4
$\varkappa_{\text{mask}}$	0,5	0,5
$\varkappa_{\text{func}}$	0,5	0,5
w_{conf}	0,5	изменяется
w_{supp}	0	0
$w_{\text{similarity}}$	0,5	изменяется

Результаты моделирования в первой серии экспериментов представлены на рис. 2.

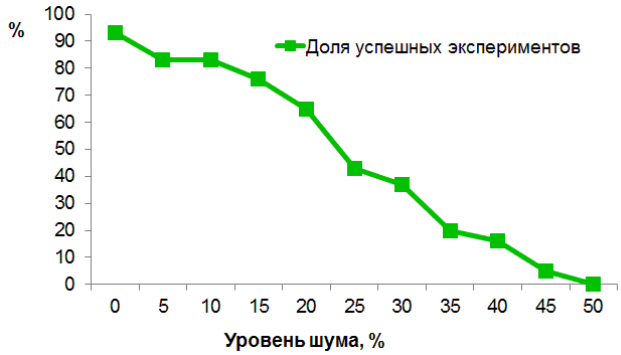


Рис. 2. Качество распознавания при различных весах функционала качества изменяющейся закономерности

Результаты показывают, что качество распознавания линейно убывает при увеличении уровня шума. При этом алгоритм поиска плавно меняющихся закономерностей является достаточно стабильным и проводит вполне эффективный интеллектуальный анализ данных даже для зашумленных пучков временных рядов.

Во второй серии экспериментов исследовалось влияние меры сходства закономерностей на качество распознавания. С этой целью проводился поиск изменяющихся закономерностей для разных комбинаций весов функционала качества Q_{step} шага изменяющейся закономерности. При этом исследование влияния меры сходства закономерностей проводилось для нескольких значений уровня шума.

Эксперименты проводились следующим образом. Был задан набор значений уровня шума ε: 0,2, 0,3 и 0,5. Для каждого значения уровня шума генерировался пучок времен-

ных рядов способом, аналогичным примененному в первой серии экспериментов. Значения параметров генерации представлены в табл. 3.

В сгенерированном пучке временных рядов происходил поиск изменяющихся закономерностей при различных весах функционала качества Q_{step} шага изменяющейся закономерности. Вес поддержки w_{supp} полагался равным нулю, что исключило влияние уровня поддержки на выбор оптимальной изменяющейся закономерности. Вес меры сходства закономерностей $w_{\text{similarity}}$ увеличивался от 0 до 1 с шагом 0,05. Вес достоверности w_{conf} соответственно уменьшался от 1 до 0 с шагом 0,05.

Для каждого значения уровня шума было сгенерировано 100 пучков временных рядов. В каждом пучке временных рядов происходил поиск плавно меняющихся закономерностей при 21-й различной комбинации весов функционала качества Q_{step}. Значения параметров алгоритмов поиска закономерностей представлены в табл. 3 и 4.

Для каждого значения уровня шума и для каждой комбинации весов была рассчитана доля успешных экспериментов. Успешный эксперимент определялся аналогично первой серии экспериментов. Доля определялась по отношению к общему количеству экспериментов для данного значения уровня шума и комбинации весов.

Результаты второй серии экспериментов представлены в табл. 5 и на рис. 3.

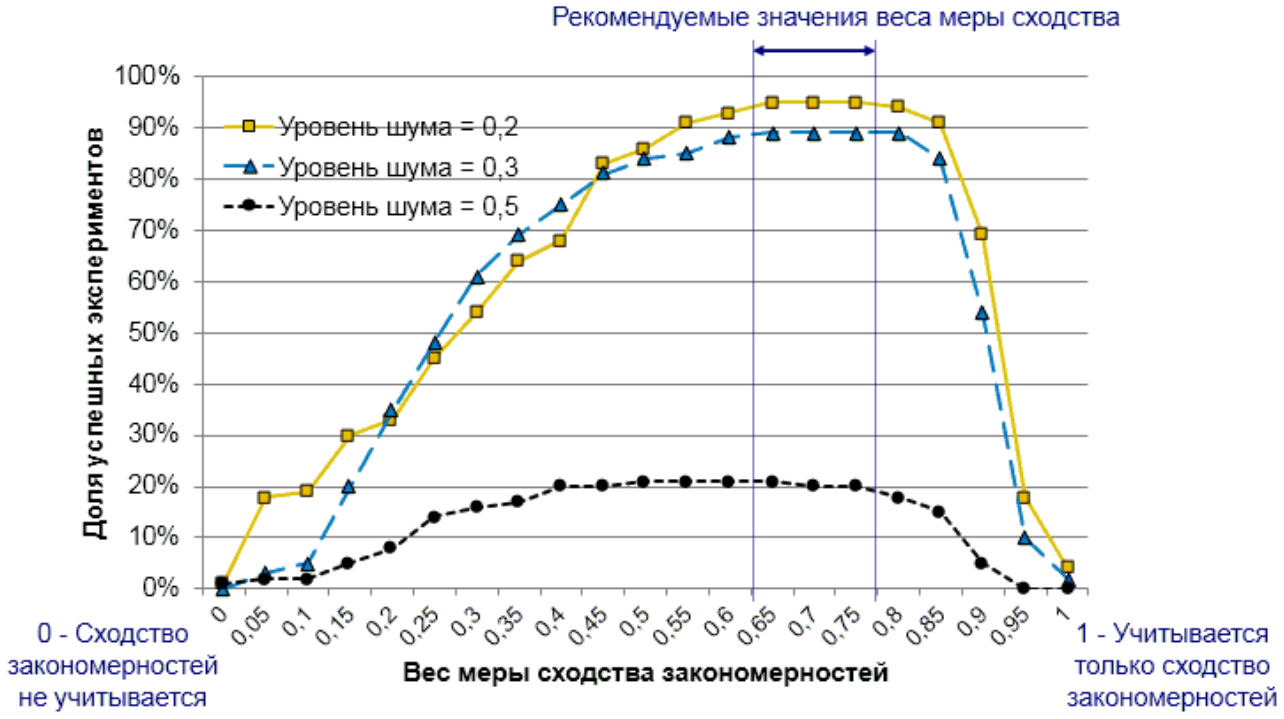


Рис. 3. Качество распознавания при различных весах функционала качества изменяющейся закономерности

Результаты второй серии экспериментов указывают на необходимость использования как меры сходства закономерностей, так и достоверности в функционале качества шага изменяющейся закономерности. При граничных значениях весов доля успешных экспериментов снижается, и, наоборот, она достигает максимального уровня при значениях веса меры сходства закономерностей $w_{\text{similarity}}$ в интервале от 0,65 до 0,8 (соответственно значениях веса достоверности w_{conf} от 0,35 до 0,2).

Комбинация значений весов $w_{\text{conf}} = 1$ и $w_{\text{similarity}} = 0$ соответствует алгоритму, при котором на каждом отрезке выбирается закономерность, обладающая максимальной достоверностью. Объединенные вместе данные закономерности составляют изменяющуюся закономерность. Мера сходства закономерностей в этом случае не используется.

При указанных значениях весов и высоком уровне шума алгоритм поиска изменяющихся закономерностей склонен к «переобучению». Так как мера сходства закономерностей не используется в функционале качества, в изменяющуюся закономерность объединяются «локальные оптимумы» каждого из отрезков. В итоге при любом уровне шума поиск плавно меняющихся закономерностей без использования меры сходства закономерностей приводит к низкому качеству распознавания: не удается правильно определить плавно меняющуюся закономерность.

Теперь рассмотрим случай, когда веса принимают другое пограничное значение: $w_{\text{conf}} = 0$ и $w_{\text{similarity}} = 1$. Тогда единственным критерием для выбора закономерностей является их близость. Если нет ограничения на количество закономерностей, записываемых в базу знаний алгоритмом поиска стационарных закономерностей, то такой выбор весов исключает возможность изменения закономерности. То есть при данном выборе весов оптимальной является любая изменяющаяся закономерность составленная из одинаковых стационарных закономерностей (т. е. фактически стационарная закономерность).

При комбинации весов $w_{\text{conf}} = 0$ и $w_{\text{similarity}} = 1$ бывает удобно упорядочить закономерности по убыванию достоверности и поставить ограничение на количество закономерностей, записываемых в базу знаний алгоритмом поиска стационарных закономерностей. Например, в базу знаний на каждом из отрезков могут записываться только 50 закономерностей, обладающих наилучшей достоверностью. Тогда достоверность будет неявно учитываться при выборе плавно меняющейся закономерности.

Таким образом, результаты второй серии экспериментов показывают, что добавление меры сходства закономерностей в функционал качества позволяет существенно повысить точность распознавания в пучках с плавно меняющимися закономерностями. При этом рекомендуемыми значениями веса меры сходства закономерностей $w_{\text{similarity}}$ являются числа в диапазоне от 0,65 до 0,8.

Примеры решения реальных задач

С целью сравнить предложенный в настоящей работе подход с другими методами была проведена серия экспериментов по краткосрочному прогнозированию временных рядов. Данными послужили курсы акций компаний Adobe, BMC, Business Objects, Cognos, Computer Associate, Novell, Oracle, Peoplesoft, Rational. Рассматривался средний почасовой курс акций в долларах за период с 13 мая 2002 г. по 10 декабря 2004 г. Средний почасовой курс получался как среднее арифметическое из четырех чисел: цены открытия (цены акции в начале часа), верхней цены (максимальной цены акции за час), нижней цены (минимальной цены акции за час), цены закрытия (цены акции в конце часа). Все девять временных рядов рассматривались как единый пучок, так как перечисленные выше компании работают в одной сфере разработки программного обеспечения для предприятий и не исключены взаимосвязи между поведением акций этих компаний.

Для прогнозирования цены акции помимо предложенного в настоящей работе метода применялось экспоненциальное сглаживание с параметром α, принимающим значения 0,1 и 0,3.

В связи с тем, что предложенный в настоящей работе подход предложен для пучков конечнозначных временных рядов, исходные действительные временные ряды были пре-

Таблица 5. Результаты экспериментов при различных весах функционала качества

w_{conf}	$w_{similarity}$	Доля успешных экспериментов, %		
		$\varepsilon = 0,2$	$\varepsilon = 0,3$	$\varepsilon = 0,5$
1,00	0,00	1	0	1
0,95	0,05	18	3	2
0,90	0,10	19	5	2
0,85	0,15	30	20	5
0,80	0,20	33	35	8
0,75	0,25	45	48	14
0,70	0,30	54	61	16
0,65	0,35	64	69	17
0,60	0,40	68	75	20
0,55	0,45	83	81	20
0,50	0,50	86	84	21
0,45	0,55	91	85	21
0,40	0,60	93	88	21
0,35	0,65	95	89	21
0,30	0,70	95	89	20
0,25	0,75	95	89	20
0,20	0,80	94	89	18
0,15	0,85	91	84	15
0,10	0,90	69	54	5
0,05	0,95	18	10	0
0,00	1,00	0	0	0

образованы в четырехзначные. Для каждого из рядов в пучке было произведено два преобразования. Первое преобразование состояло в переходе от исходных значений к разностям. Второе сопоставило каждой разности элемент алфавита $E_4 = \{0, 1, 2, 3\}$. Группировка действительных значений осуществлялась разбиением на квартили: 25 процентам разностей ставится в соответствие 0, следующим 25 процентам разностей ставится в соответствие 1 и т. д. Разбиение на квартили для каждого из рядов происходило независимо.

При прогнозировании действительных временных рядов с использованием предложенного метода производились обратные преобразования. Выбирался последний шаг плавно меняющейся закономерности — стационарная закономерность последнего отрезка и применялась к известной части пучка временных рядов. На основе прогнозируемого значения из E_4 определялось прогнозируемое изменение цены акции, которое добавлялось к последнему известному значению исходного ряда. Таким образом получался прогноз на 1 шаг вперед для целевого ряда в исходном пучке временных рядов.

Соответствие, полученное в результате дискретизации, приводится в табл. 6 и 7. Для каждого временного ряда в таблицах представлены диапазоны разностей, которым ставится в соответствие значение из E_4. Отдельно выделено среднее значение в каждом квартиле. Оно используется при обратном переходе от конечнозначных временных рядов к действительным с целью определения прогнозируемого значения исходно ряда.

Таблица 6. Результаты дискретизации

Ряд	От	До	Значение из E_4	Среднее
Adobe	$-\infty$	−0,11750	0	−0,29070
Adobe	−0,11750	0,00650	1	−0,05033
Adobe	0,00650	0,12750	2	0,06096
Adobe	0,12750	∞	3	0,29942
BMC	$-\infty$	−0,05500	0	−0,14016
BMC	−0,05500	0,00000	1	−0,02573
BMC	0,00000	0,05250	2	0,02359
BMC	0,05250	∞	3	0,14402
Business Objects	$-\infty$	−0,09500	0	−0,27475
Business Objects	−0,09500	0,00000	1	−0,04382
Business Objects	0,00000	0,09250	2	0,04332
Business Objects	0,09250	∞	3	0,26859
Cognos	$-\infty$	−0,09250	0	−0,21977
Cognos	−0,09250	0,00250	1	−0,04161
Cognos	0,00250	0,09500	2	0,04365
Cognos	0,09500	∞	3	0,23489
Computer Associate	$-\infty$	−0,06500	0	−0,16745
Computer Associate	−0,06500	0,00250	1	−0,02772
Computer Associate	0,00250	0,07250	2	0,03438
Computer Associate	0,07250	∞	3	0,17251

Модели экспоненциального сглаживания получали на вход исходные действительные временные ряды, что позволило этим методам использовать всю доступную информацию.

При сравнении предложенного метода и экспоненциального сглаживания каждый из методов осуществил 20 прогнозов на один момент времени вперед. Средний квадрат ошибки каждого из методов представлен в табл. 8.

Как видно из табл. 8, при прогнозировании курса акций предложенный метод превосходит по качеству прогнозирования экспоненциальное сглаживание при некоторых параметрах.

Помимо прогнозирования значений пучка временных рядов предложенный в работе алгоритм поиска плавно изменяющихся закономерностей позволил получить представление о характере структурных изменений в пучках временных рядов.

Например, для целевого ряда, описывающего поведение курса акций компании Rational, была найдена следующая плавно меняющаяся закономерность. Шаги плавно меняющейся закономерности — это стационарные закономерности для трех отрезков пучка. Первый отрезок начинается датой 13 мая 2002 г., второй — 21 февраля 2003 г., третий — 28 ноября 2003 г.

Приведенную закономерность можно описать и в терминах действительных временных рядов с помощью табл. 6 и 7. Например, первый столбец закономерности может быть интерпретирован следующим образом: если на последнем временном интервале (час) курса акций Rational упал более чем на 0,0675 доллара США и курс акций Computer Associates

Таблица 7. Результаты дискретизации

Ряд	От	До	Значение	Прогноз
Novell	$-\infty$	−0,02500	0	−0,07158
Novell	−0,02500	−0,00175	1	−0,01273
Novell	−0,00175	0,02500	2	0,01024
Novell	0,02500	∞	3	0,07677
Oracle	$-\infty$	−0,04000	0	−0,09399
Oracle	−0,04000	−0,00025	1	−0,01922
Oracle	−0,00025	0,03750	2	0,01709
Oracle	0,03750	∞	3	0,10058
Peoplesoft	$-\infty$	−0,07000	0	−0,18676
Peoplesoft	−0,07000	−0,00175	1	−0,03260
Peoplesoft	−0,00175	0,06500	2	0,02850
Peoplesoft	0,06500	∞	3	0,19325
Rational	$-\infty$	−0,06750	0	−0,20687
Rational	−0,06750	0,00500	1	−0,02739
Rational	0,00500	0,07500	2	0,03584
Rational	0,07500	∞	3	0,19421

Таблица 8. Средний квадрат ошибки

Целевой Предложенный ряд	Экспоненциальное сглаживание $\alpha = 0{,}1$	Экспоненциальное сглаживание $\alpha = 0{,}3$	метод
Adobe	$9{,}05 \cdot 10^{-2}$	$7{,}32 \cdot 10^{-2}$	$6{,}89 \cdot 10^{-2}$
BMC	$13{,}24 \cdot 10^{-3}$	$11{,}15 \cdot 10^{-3}$	$9{,}72 \cdot 10^{-3}$
Business Objects	$17{,}42 \cdot 10^{-2}$	$7{,}19 \cdot 10^{-2}$	$3{,}74 \cdot 10^{-2}$
Cognos	$6{,}73 \cdot 10^{-2}$	$3{,}08 \cdot 10^{-2}$	$2{,}39 \cdot 10^{-2}$
Computer Associates	$4{,}49 \cdot 10^{-2}$	$2{,}87 \cdot 10^{-2}$	$1{,}91 \cdot 10^{-2}$
Novell	$7{,}62 \cdot 10^{-3}$	$4{,}18 \cdot 10^{-3}$	$2{,}54 \cdot 10^{-3}$
Oracle	$8{,}61 \cdot 10^{-3}$	$6{,}77 \cdot 10^{-3}$	$5{,}45 \cdot 10^{-3}$
Peoplesoft	$3{,}24 \cdot 10^{-3}$	$2{,}94 \cdot 10^{-3}$	$1{,}26 \cdot 10^{-3}$
Rational	$5{,}19 \cdot 10^{-2}$	$3{,}43 \cdot 10^{-2}$	$2{,}06 \cdot 10^{-2}$

Таблица 9. Первый шаг плавно меняющейся закономерности ряда Rational

Rational (t-1)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Computer Associates (t-8)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Rational (t)	0	0	0	0	0	0	0	0	3	0	3	3	3	3	3	3

уменьшился более чем на 0,065 доллара, то курс акций Rational на следующем временном интервале упадет примерно на 0,2 доллара.

Таблица 10. Второй шаг плавно меняющейся закономерности ряда Rational

Rational (t-1)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Computer Associates (t-**6**)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Rational (t)	0	0	0	0	**1**	**1**	0	**1**	**2**	**1**	**2**	**2**	3	3	3	**2**

Жирным шрифтом выделены изменения при переходе к следующему шагу плавно меняющейся закономерности.

Таблица 11. Третий шаг плавно меняющейся закономерности ряда Rational

Rational (t-1)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Computer Associates (t-6)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Rational (t)	0	0	**1**	0	1	1	**1**	1	**1**	**2**	2	2	3	3	**2**	**3**

При переходе от первого шага ко второму произошли более существенные структурные изменения закономерности по сравнению с переходом от второго шага к третьему. Последующий анализ событий на рынке акций показал, что указанное структурное изменение могло быть вызвано покупкой компанией IBM компании Rational, произошедшей 20 февраля 2003 г. Данное событие повлияло на структуру компании Rational и на отношение инвесторов к данной компании. В свою очередь указанные изменения отразились на закономерностях, определяющих поведение временного ряда курса акций компании Rational.

Таким образом, технический анализ пучка временных рядов с использование алгоритма поиска плавно меняющихся закономерностей позволяет выявлять события, которые влияют на закономерности, определяющие поведение рядов.

Результаты испытаний показали, что предложенный в настоящей работе подход может быть более эффективен при краткосрочном прогнозировании, чем другие алгоритмы прогнозирования. При этом алгоритм не только позволяет сделать прогноз, но и осуществляет поиск скрытых закономерностей, описывающих явление.

Заключение

В настоящей работе рассматривается подход к поиску закономерностей в пучках конечнозначных временных рядов. Этот подход позволяет выявлять закономерности, которые подвергаются «плавным» структурным изменениям с течением времени. Для определения подобного рода изменений в работе описана мера сходства закономерностей и описано ее применение как одного из весов на графе закономерностей.

Разрабатываемый алгоритм поиска плавно меняющихся закономерностей в пучках временных рядов решает задачу как алгоритм интеллектуального анализа данных. Он не только позволяет прогнозировать процесс, но и осуществляет поиск скрытых закономерностей в данных и дает возможность в явном виде описать закономерность. Найденные закономерности могут быть использованы как для прогнозирования следующих элементов пучка временных рядов, так и для детального анализа явления, описанного пучком временных рядов, и моделирования явления. Это делает возможным применение предло-

женного алгоритма в широком пласте задач прогнозирования временных рядов, а также в задачах изучения и описания процессов, которые могут представлены пучком временных рядов.

Предложенный в настоящей работе подход был реализован в программной системе и протестирован на модельных и реальных задачах. Испытания на модельных задачах с использованием разработанного экспериментального стенда показали, что алгоритм, основанный на введенных в работе мерах сходства и функционалах качества, позволяет эффективно находить заложенные закономерности, в том числе при достаточно высоком уровне шума.

Эксперименты на модельных пучках временных рядов показали, что использование введенной меры сходства закономерностей в функционале качества существенно повышает качество прогнозирования. Вместе с тем был получен диапазон весов, при котором достигается максимальное качество распознавания.

Анализ реальных временных рядов с применением предложенного алгоритма также свидетельствовал об эффективности алгоритма при краткосрочном прогнозировании. Вместе с тем алгоритм решает и задачу интеллектуального анализа данных, предложив закономерности, описывающие взаимосвязь одномерных временных рядов.

Таким образом, апробация предложенного подхода к прогнозированию процессов с плавно меняющимися закономерностями на модельных и реальных данных позволяет судить о достаточной эффективности предложенных алгоритмов при анализе пучков временных рядов с плавно меняющимися закономерностями.

Литература

[1] *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.

[2] *Бокс Дж., Дженкинс Г.* Анализ временных рядов, прогноз и управление. М.: Мир, 1974.

[3] *Хеннан Э.* Многомерные временные ряды. М.: Мир, 1974.

[4] *Engle R. F., Kroner K. F.* Multivariate Simultaneous Generalized ARCH // *Econometric Theory*, 1993. Vol. 11, P. 122–150.

[5] *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003. 416 с.

[6] *Agrawal R., Imielinski T., Swamiet A.* Mining association rules between sets of items in large databases // *Conference Management of Data Proceedings*, 1993. P. 207–216.

[7] *Agrawal R., Srikant R.* Mining sequential patterns // *11th Conference (International) on Data Engineering Proceedings*, 1995. P. 3–14.

[8] *Mannila H., Toivonen H., Verkamo A. I.* Discovery of frequent episodes in event sequences // *Data Mining Knowledge Discovery*, 1997. Vol. 1, No. 3. P. 259–289.

[9] *Morchen F., Ultsch A.* Efficient mining of understandable patterns from multivariate interval time series // *Data Mining Knowledge Discovery*, 2007. Vol. 15, No. 2. P. 181–215.

[10] *Das G., Lin K., Mannila H., et al.* Rule discovery from time series // *4th Conference (International) on Knowledge Discovery and Data Mining Proceedings*, 1998. P. 16–22.

[11] *Sayal M.* Detecting time correlations in time-series data streams. Palo Alto: HP Labs, 2004.

[12] *Morchen F., Ultsch A.* Optimizing time series discretization for knowledge discovery // *11th Conference (International) on Knowledge Discovery and Data Mining Proceedings*, 2005. P. 660–665.

[13] *Brown R. G.* Smoothing forecasting and prediction of discrete time series. N.Y.: Prentice-Hall, 1963.

[14] *Trigg D. W., Leach A. G.* Exponential smoothing with an adaptive response rate // *Operat. Res. Quart.*, 1967. Vol. 18, No. 1. P. 53–59.

[15] *Engle R. F.* ARCH: Selected readings. Oxford: Oxford Univ. Press, 1995.

[16] *Zadeh L. A., Ragazzini J. R.* The analysis of sampled-data systems // *Applic. Industry (AIEE)*, 1952. P. 225–234.

[17] *Rao A. G., Shapiro A.* Adaptive smoothing using evolutionary spectra // *Management Sc.*, 1970. Vol. 17, No. 3. P. 208–218.

[18] *Filipenkov N. V.* Data mining in non-stationary multidimensional time series using a rule similarity measure // *IADIS European Conference on Data Mining Proceedings*, 2008. P. 92–96.

[19] *Филипенков Н. В.* Об одном методе поиска плавно меняющихся закономерностей в пучках временных рядов // *Ж. вычисл. матем. и матем. физики*, 2009. Т. 49, № 11. С. 2020–2040.

[20] *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004.

[21] *Журавлев Ю. И.* Избранные научные труды. М.: Магистр, 1998.

[22] *Рудаков К. В.* Алгебраическая теория универсальных и локальных ограничений для алгоритмов распознавания. Дисс. ... докт. физ.-мат. наук. М.: ВЦ РАН, 1992.

[23] *Кристофидес Н.* Теория графов. Алгоритмический подход. М.: Мир, 1978. 432 с.

References

[1] *Anderson T. W.* 1971. The Statistical analysis of time series. N.Y: John Wiley & Sons.

[2] *Box G., Jenkins G.* 1970. Time series analysis: Forecasting and control. San Francisco: Holden-Day.

[3] *Hannan E. G.* 1970. Multiple time series. N.Y: John Wiley & Sons.

[4] *Engle R. F., Kroner K. F.* 1993. Multivariate Simultaneous Generalized ARCH. *Econometric Theory* 11:122–150.

[5] *Lukashin Yu. P.* 2003. Adaptive methods for short-term forecasting of Time Series. Moscow: Finansy i Statistika. 416 p. (in Russ.)

[6] *Agrawal R., Imielinski T., Swamiet A.* 1993. Mining association rules between sets of items in large databases. *Conference Management of Data Proceedings* 207–216.

[7] *Agrawal R., Srikant R.* 1995. Mining sequential patterns. *11th Conference (International) on Data Engineering Proceedings* 3–14.

[8] *Mannila H., Toivonen H., Verkamo A. I.* 1997. Discovery of frequent episodes in event sequences. *Data Mining Knowledge Discovery* 1(3):259–289.

[9] *Morchen F., Ultsch A.* 2007. Efficient mining of understandable patterns from multivariate interval time series. *Data Mining Knowledge Discovery* 15(2):181–215.

[10] *Das G., Lin K., Mannila H., et al.* 1998. Rule discovery from time series. *4th Conference (International) on Knowledge Discovery and Data Mining Proceedings* 16–22.

[11] *Sayal M.* 2004. Detecting time correlations in time-series data streams. Palo Alto: HP Labs.

[12] *Morchen F., Ultsch A.* 2005. Optimizing time series discretization for knowledge discovery. *11th Conference (International) on Knowledge Discovery and Data Mining Proceedings* 660–665.

[13] *Brown R. G.* 1963. Smoothing forecasting and prediction of discrete time series. New York: Prentice-Hall.

[14] *Trigg D. W., Leach A. G.* 1967. Exponential smoothing with an adaptive response rate. *Operat. Res. Quart.* 18(1):53–59.

[15] *Engle R. F.* 1995. ARCH: Selected readings. Oxford: Oxford Univ. Press.

[16] *Zadeh L. A., Ragazzini J. R.* 1952. The analysis of sampled-data systems. *Applic. Industry (AIEE)* 225–234.

[17] *Rao A. G., Shapiro A.* 1970. Adaptive smoothing using evolutionary spectra. *Management Sc.* 17(3): 208–218.

[18] *Filipenkov N. V.* 2008. Data mining in non-stationary multidimensional time series using a rule similarity measure. *IADIS European Conference on Data Mining Proceedings* 92–96.

[19] *Filipenkov N. V.* 2009. A method for finding smoothly varying rules in multidimensional time series. *Computational Mathematics Mathematical Physics* 49(11):1930–1948.

[20] *Barsegyan A. A., Kupriyanov M. S., Stepanenko V. V., Kholod I. I.* 2004. Methods and models of data analysis: OLAP and Data Mining. St. Petersburg: BKhV-Peterburg. (in Russ.)

[21] *Zhuravlev Yu. I.* 1998. Selected scientific works. Moscow: Magistr. (in Russ.)

[22] *Rudakov K. V.* 1992. D.Sc. Diss. Moscow: Dorodnicyn Computing Centre of the Russian Academy of Sciences. (in Russ.)

[23] *Christofides N.* 1975. Graph theory: An algorithmic approach. Orlando, FL: Academic Press. 400 p.

Определение местоположения телефона по данным сенсоров*

А. А. Остапец
aostapec@mail.ru
Факультет вычислительной математики и кибернетики МГУ, Москва, Ленинские горы, МГУ, 2-й учебный корпус

Данная статья посвящена использованию методов машинного обучения в задаче определения местоположения телефона (сумка, карман, рука), который несет движущийся человек. Задача является актуальной и имеет множество практических применений, как, например, автоматическое включение/выключение энергозатратных сервисов при различном положении мобильного устройства. Поставленная задача решается по сигналам двух датчиков телефона – акселерометра и гироскопа. Основной смысл работы – это способ выбора и предобработки признаков, позволяющий уменьшить влияние шума на результат классификации и анализировать активность в независимости от пространственной ориентации мобильного устройства. Результаты, полученные в ходе вычислительного эксперимента, подтверждают применимость предложенного подхода.

Ключевые слова: обработка сигналов; сенсоры; акселерометр; гироскоп; машинное обучение.

Smartphone location recognition using mobile sensors*

A. A. Ostapets
MSU Faculty of Computational Mathematics and Cybernetics, Moscow

This article focuses on the use of machine learning methods in the task of determining the location of the phone (bag, pocket, hand). This problem is important in many practical applications, such as automatic on / off energy-intensive services at various positions of the mobile device. The aim of this study was to evaluate and validate the possibility of detecting mobile phone place. The data were collected using the accelerometer and the gyroscope. The whole classification process (preprocessing, feature extraction and classification) is presented in this article. Acceleration data acquired suffers from changes due accelerometer noise which needs to be eliminated. It's solved using low-pass filter. Primary features that are often used when working with the signals from the sensors are described in this paper. Feature selection was conducted on real data, and the best features were selected. Algorithms have trained using phone orientation independent features to recognize several locations of the phone. It's was tested by two different datasets. The paper presents an experimental study and comparative analysis of algorithms. It is shown that the proposed approach achieved 88% accuracy on the used datasets.

Keywords: signal processing; sensors; accelerometer; gyroscope; machine learning.

Введение

Существуют много публикаций, которые посвящены задаче классификации вида физической активности человека и идентификации по походке. Например в статье [1] рас-

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-07-00965.

сматривается задача определения активности человека по данным пяти акселерометров, закрепленных на человеке. В статье [2] авторы решают проблему классификации шагов пешехода. Они выделяют три класса: шаги по ровной поверхности, шаги по лестнице вверх или вниз. В статье [3] осуществляется попытка предсказать текущую частоту сердечного ритма человека по данным от сенсоров смартфона.

Задача определения местоположения произвольного телефона для любого пользователя является сложной по следующим причинам: манера движения, в частности походка, у людей сильно различается; характеристики одежды, карманов и сумок варьируются в широких пределах, ориентация телефона в пространстве может быть произвольной. Датчики мобильных устройств имеют значительный разброс параметров.

В статье [4] рассматривается вопрос определения верной позиции смартфона. Авторы выделяют девять возможных позиций, в которых может находиться телефон. Для классификации используются алгоритм SVM, который обучают на 60 отобранных признаках. Эксперимент производился по методике Leave-One-Out: на всех пользователях, кроме одного, производится обучение, на отобранном пользователе тестирование. Итоговая точность по всем классам получается около 75%.

В статье [5] рассматривается вопрос определения верной позиции сенсора на теле человека. Сенсор может находиться в одной из шести различных позиций. Стоит отметить, что в данной работе использовались медицинские датчики, которые отличаются повышенной точностью. Для классификации также использовался алгоритм SVM. Эксперимент производился по следующей методике: здесь в обучение могли попадать временные ряды того человека, на котором производится в данный момент классификация. Поэтому итоговая точность в данном случае получается намного выше – около 89%.

Сенсоры

Сенсором называется устройство, преобразующее измеряемую величину в сигнал для последующего анализа. В данной работе рассматривается работа с двумя сенсорами: акселерометром и гироскопом.

Основным назначением акселерометра является предоставление информации о текущем ускорении устройства, вернее разности ускорения устройства и ускорения свободного падения. В состоянии покоя показания датчика совпадают с вектором ускорения свободного падения. В условиях невесомости истинное ускорение объекта вызывается лишь гравитационной силой и потому в точности равно гравитационному ускорению. Таким образом, кажущееся ускорение отсутствует и показания любого акселерометра равны нулю [6].

Гироскоп измеряет угловую скорость. Обычно используется совместно с акселерометром для отслеживания изменений в движениях. В электронных устройствах программное обеспечение, используемое вместе с гироскопом, способно быстро реагировать на перемещение устройства в пространстве и принимать соответствующие решения. Например, в ноутбуках гироскоп позволяет быстро включить режим фиксации жесткого диска в случае падения или просто резкого перемещения устройства. В мобильных устройствах используются датчики угловой скорости. Этот тип гироскопов является намного более простым и дешёвым при достаточно высокой точности [7].

Описание данных

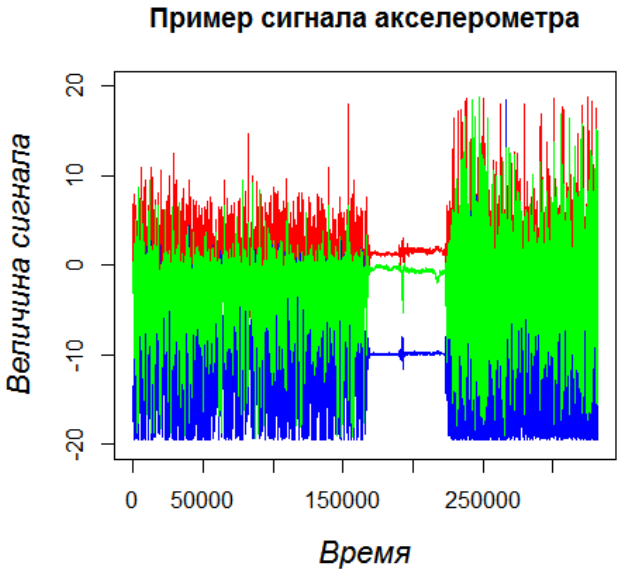
В рамках данной работы ставилась задача исследования возможности определения местоположения мобильного устройства с помощью данных акселерометра и гироскопа. Было выделено 3 основных месторасположения телефона: сумка(bag), карман(pocket),

рука(hand). Каких-либо ограничений на ориентацию устройства в пространстве не ставилось – в любом месторасположении телефон может находиться в произвольной ориентации.

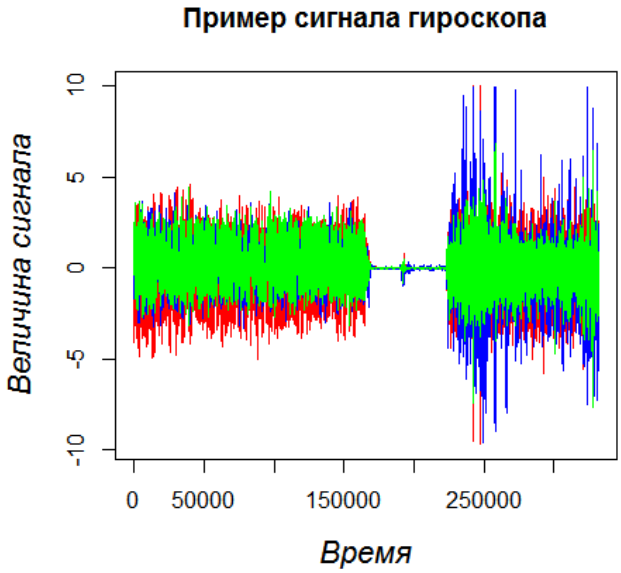
Данные, на которых проводились эксперименты:

— Собственный набор данных. Эти данные содержат результаты измерений для 9 человек. Все люди осуществляли одинаковые действия - ходили по одной лестнице вверх и вниз. Частота дискретизации сигналов для всех наборов данных равна 100 Гц. Практически для каждого человека присутствуют 3 различных месторасположения телефона. Информация собиралась с помощью разных устройств (Nokia Lumia 720, Samsung Galaxy, ...). Для некоторых данных (например, для тех которые получены с помощью Lumia 720), отсутствуют показания гироскопа.

— Общедоступный набор данных Walk Detection and Step Counting on Unconstrained Smartphones, который содержит измерения для 27 человек. Каждый участник проходил одну и ту же дистанцию. Данные собирались с помощью телефона Galaxy Nexus GT-I9250 под системой Android 4.1.1, с помощью акселерометра Bosch BMA220 с частотой данных 100 Гц). Подробное описание данных содержится в статье [8]. Ниже приведена статистика о пользователях из этого набора данных.



(а) Пример данных акселерометра



(б) Пример данных гироскопа

Таблица 1. Пол

Мужчин	18
Женщин	9

Таблица 2. Возраст

15-19	9
20-29	18

Таблица 3. Рост[см]

150-159	3
160-169	5
170-179	11
180-189	8

Фильтр нижних частот

Показания акселерометра и гироскопа на мобильных устройствах подвержены достаточно сильному шуму. Например, погрешность показаний акселерометра иногда достигает 0.05g, поэтому необходима борьба с шумом. Фильтры нижних частот – это группа фильтров основной особенностью которых является способность фильтровать сигналы выше указанной частоты, то есть такие фильтры пропускают сигналы низкой частоты, что позволяет избавиться от шумовых помех сигнала.

Самый простой фильтр нижних частот описывается следующей формулой:

$$O_n = O_{n-1} + \alpha(I_n - O_{n-1}),$$

где O_n – выходное значение сигнала (отфильтрованное), I_n – входное значение (нефильтрованное), α – коэффициент фильтрации, принимающий значения от 0 до 1. При α равном 1, выходные значения совпадают с входными. В качестве примера, на Рис. 1 приводятся два варианта фильтрованных с помощью фильтра нижних частот данных с коэффициентом $\alpha = 0.2$ и $\alpha = 0.05$, соответственно.

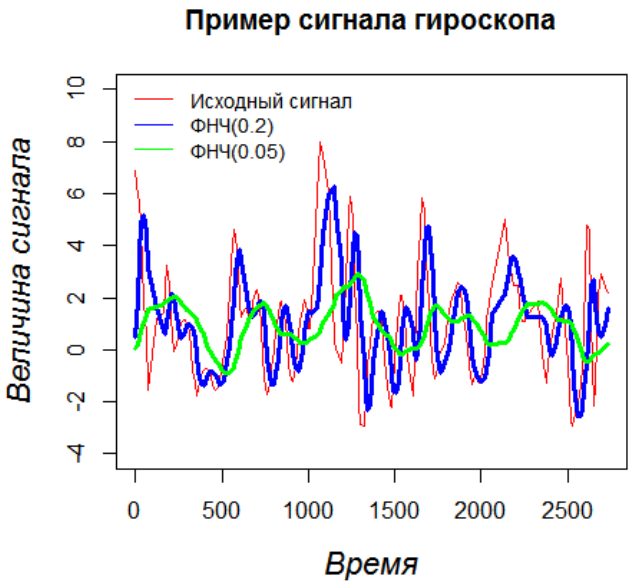


Рис. 1. Применение фильтра нижних частот для фильтрации сигнала

Как видно из примеров меньший коэффициент дает более гладкий результат. Получающийся в результате сигнал достаточно гладкий, но, так же как и при сглаживании методом скользящего среднего, присутствует некоторая задержка, особенно при резком колебании значений. Фильтр нижних частот подавляет сигналы выше некоторой критической частоты w и пропускает сигналы ниже этой частоты. Пример применения данного метода для работы с сигналами от акселерометра в статьях [9], [10].

Дополнительная очистка сигнала

Помимо погрешностей в показаниях сенсоров существует проблема с неоднородностью деятельности человека при сборе данных. Например, на Рис. 2 представлены показания акселерометра при ходьбе человека по лестнице вверх и вниз. Участок в середине наблюдения соответствует остановке и смене направления движения. Для многих задач

распознавания такой участок будет только мешать решению, особенно, когда обучение и классификация происходит на небольших участках сигнала – окнах.

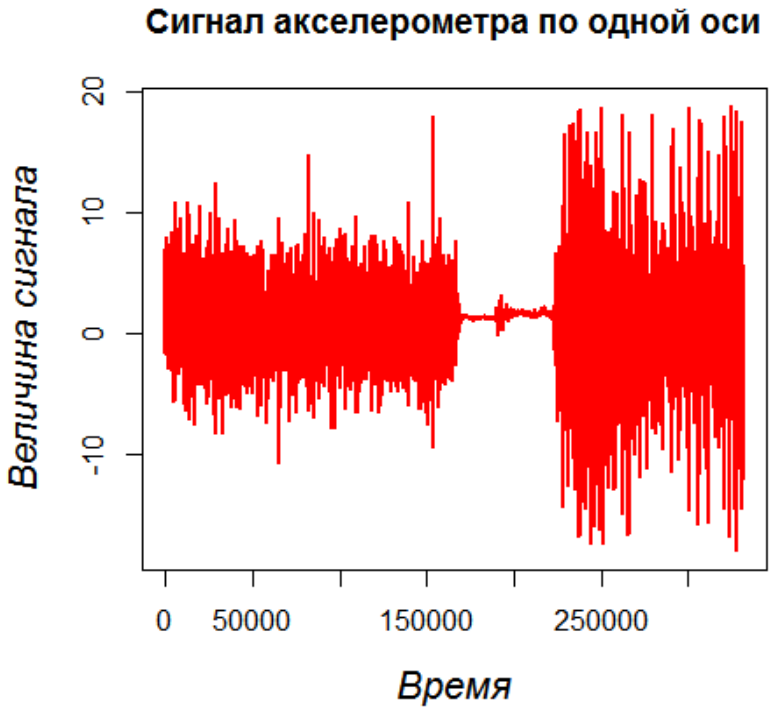


Рис. 2. Пример сигнала акселерометра за длительный промежуток времени

В рамках данной задачи было принято решение удалять такие участки.

Выделение признаков

Даже при небольшом изменении ориентации телефона в пространстве сигналы получаемые от сенсоров могут «значимо» измениться. Под термином «значимо» здесь подразумевается неинвариантность к изменению направления телефона основных статистических характеристик, таких как среднее значение, дисперсия, квантили, . . . на каждую ось сенсора.

В данной работе исследуется вопрос определения местоположения телефона при отсутствии каких-либо ограничений на ориентацию телефона в пространстве. Поэтому особенно важно найти такие признаки, которые не зависят или хотя бы минимизируют влияние направления телефона.

Проекции на вертикальную ось и горизонтальную плоскость

Для уменьшения влияния ориентации телефона разложим исходный сигнал акселерометра на две составляющих – вертикальную и горизонтальную компоненты [11].

Оценка вектора силы тяжести может быть получена путем усреднения показаний акселерометра. Это оценка вектора гравитации в свою очередь, позволяет оценить вертикальную составляющую и величину горизонтальной составляющей движения пользователя, независимо от того, как ориентированы оси акселерометра. В идеале, хотелось бы извлечь информацию об ускорении с точки зрения системы координат, сопоставленной с движением человека. Для наблюдений акселерометра получаем оценку вектора силы тяжести

на каждую ось путем усреднения всех показаний акселерометра на эту ось. Фактически, мы оцениваем вертикальный вектор ускорения v соответствующей вектору силы тяжести как $v = (v_x, v_y, v_z)$, где v_x, v_y и v_z являются средними значениями всех измерений интервала дискретизации на соответствующие оси.

Обозначим через $\mathbf{a} = (a_x, a_y, a_z)$ вектор, который является одним измерением внутри окна. Тогда через $\mathbf{d} = (a_x - v_x, a_y - v_y, a_z - v_z)$ обозначим динамическую компоненту, которая вызвана движения пользователя, а не гравитацией. Затем, с помощью скалярного произведения, можно вычислить проекцию $\mathbf{p}$ вектора $\mathbf{d}$ на вертикальную ось $\mathbf{v} : \mathbf{p} = (\frac{\mathbf{d}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}})\mathbf{v}$.

Далее, так как вектор является суммой своих вертикальной и горизонтальной компонент, можно вычислить горизонтальную составляющую динамического ускорения путем вычитания векторов $\mathbf{h} = \mathbf{d} - \mathbf{p}$. Однако, в отличие от вертикального случая, мы не знаем ориентацию горизонтальной оси $\mathbf{h}$ относительно оси $\mathbf{f}$. На самом деле ее и невозможно обнаружить. Там нет статического ускорения, которое присутствует в вертикальном случае. Соответственно, можно вычислить общее ускорение на горизонтальную плоскость составляющей динамического ускорения, поскольку это лучшее, что можно сделать в данном случае.

Фильтрация ударов

Основной проблемой при работе с сигналами акселерометра является отсутствие информации об ориентации осей акселерометра. Поэтому нельзя отделить ускорение, вызванное деятельностью человека от ускорения силы тяжести и определить направления наблюдаемых ускорений точно. В качестве одного из решений рассматривается переход к новому каналу, где анализируется изменение ускорений вместо оригинального сигнала ускорения. Общая величина изменения ускорения является полностью независимой от ориентации и отражает ускорение, которое связано только с движением человека. Если направление гравитации может быть аппроксимировано хотя бы приблизительно, то такой канал может дать ценную информацию о изменениях направления движения.

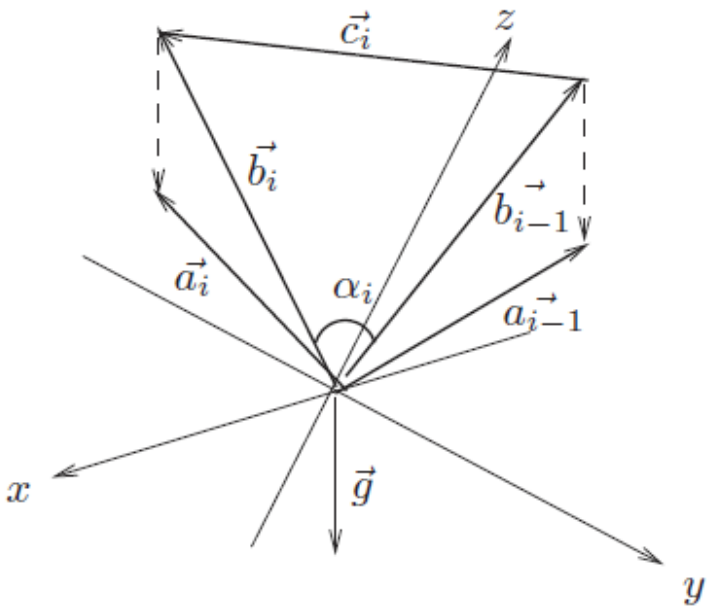


Рис. 3. Разложение сигнала на компоненты

В идеальном случае все измерение ускорения акселерометром $\boldsymbol{a}$ зависят от двух компонент: гравитационного ускорения $\boldsymbol{g}$ и ускорения $\boldsymbol{b}$, которое связано с движением человека: $\boldsymbol{a} = \boldsymbol{g} + \boldsymbol{b}$. Идея фильтрации ударов состоит в рассмотрении рывков (изменений ускорения) вместо самого ускорения. Этот подход дает возможность получить признаки, которые не зависят от ориентации телефона без необходимости оценки вектора гравитации точно.

Базовая идея метода показана на Рис. 3. Из двух последовательных ускорений $\boldsymbol{a}_{i-1}$ и $\boldsymbol{a}_i$ мы составляем вектор их разности $\boldsymbol{c}_i = \boldsymbol{a}_{i-1} - \boldsymbol{a}_i$. Это соответствует среднему рывку на временном интервале $[t_{i-1}, t_i]$. Предполагая, что ориентация телефона телефона не изменилась на этом временном отрезке, получаем, что гравитационная компонента одинакова на этих двух временных шагах, т.е $\boldsymbol{g}_{i-1} = \boldsymbol{g}_i$. Тогда $\boldsymbol{c}_i = \boldsymbol{b}_i - \boldsymbol{b}_{i-1}$ и получаем разницу без учета вектора $\boldsymbol{g}$. Дополнительно получаем независимость от ориентации телефона величины ускорения вектора $\boldsymbol{c}_i$.

Обозначим угол между векторами $\boldsymbol{b}_i$ и $\boldsymbol{b}_{i-1}$ за $\boldsymbol{\alpha}_i$. Этот угол также не зависит от ориентации телефона, но точность его оценки сильно зависит от точности оценки вектора $\boldsymbol{g}$.

Одним из примеров применения фильтрации ударов является фильтр, приведенный в работе [12]. Он определяется следующей функцией:

$$f(c_i, \alpha_i) = \Big(1 + \frac{|\alpha_i|}{180}\Big) c'_i,$$

где

$$c'_i = \begin{cases} |\boldsymbol{c}_i|, & \text{если } |\boldsymbol{a}_i| \geqslant |\boldsymbol{a}_{i-1}|, \\ -|\boldsymbol{c}_i|, & \text{иначе.} \end{cases}$$

Здесь c'_i является величиной вектора изменения ускорения, знак которого показывает увеличилось ускорение или уменьшилось. Угол α_i измеряется в градусах и принимает значения в интервале [-180, 180]. Поскольку оценка углов изменения направления является менее точной, чем величина рывка, то оценка угла используется только для модификации. Если направление не изменилось, то $f(c_i, \alpha_i)$ является просто разностью векторов $\boldsymbol{b}_i - \boldsymbol{b}_{i-1}$. Чем больше изменяется направление, тем больший коэффициент стоит при этой разности.

Признаки от сигналов акселерометра

Для работы с данными от акселерометра было выбрано 6 групп признаков:

1. Первая группа признаков, которая включает в себя стандартные статистические характеристики сигнала (min, max, mean, ...).
2. Вторая группа – это часто использующиеся признаки при работе с сигналами (SMA, MMV, AAD).
3. Третья группа – это признаки, заимствованные из анализа EEG сигналов. Впервые были представлены в статье [13].
4. Четвертая группа признаков основана на использовании числа пересечений сигналом определенных значений(crossing rate, cr). Вычисляется эта характеристика для произвольного сигнала {s(t), t = 0, ..., T} следующим образом:

$$\mathrm{cr} = \sum_{t=1}^{T-1} \mathbb{I}\left\{s'_t s'_{t-1} < 0\right\},$$

где $s'(t)$ – это сигнал, полученный после вычитания из исходного сигнала значения соответствующего квантиля.

5. Пятая группа – это признаки, состоящие из различных коэффициентов авторегрессии. К ним относятся статистика Дарбина-Уотсона (Durbin-Watson autocorrelation) [14], максимальное значение корреляции, период автокорреляции [15].
6. Шестая группа признаков включает себя корреляции между различными каналами.

Признаки от сигналов гироскопа

Для гироскопа были выбраны следующие признаки:

1. Средние значения для каждой из осей: $\overline{x}, \overline{y}, \overline{z}$.
2. Стандартные отклонения для каждой из осей: s_x^2, s_y^2, s_z^2.
3. Корреляция между показаниями различных осей: r_{XY}, r_{XZ}, r_{YZ}.
4. Площадь под величиной сигнала $= \frac{1}{T}\sum_{i=1}^{T} |x_i| + |y_i| + |z_i|$.
5. Мощность сигнала по каждой из осей: $P_v = \frac{1}{T}\sum_{i=1}^{T} v_i^2$.
6. Среднее значение величины сигнала $= \sqrt{\overline{x}^2 + \overline{y}^2 + \overline{z}^2}$.

Вычислительные эксперименты

При решении задачи важно, чтобы построенный алгоритм был независим от конкретного пользователя. Таким образом, не нужно каждый раз проводить предварительную настройку на пользователя, а можно сразу проводить классификацию на новых данных. Поэтому все эксперименты осуществлялись по методике leave-one-out: контроль – это все данные одного человека, обучение – данные от всех других людей.

Для работы с сигналами была выбрана техника 50% пересекающих окон [16]. Два соседних окна имеют пересечение в 256 измерений. Для обучения выбирается окно небольшого размера (для каждого сигнала рассматриваются окна по 512 отсчетов). Для этого окна мы знаем верный класс и можем извлечь признаки. Таким образом формируется обучающая выборка. В контроле считаем, что нужно определить месторасположение телефона по 5 окнам, которые пересекаются на 50%, т.е информации за 1536 измерений. В качестве итогового ответа выбирался самый часто предсказываемый класс: $\arg\max\limits_i \sum_{w=1}^{5} [p_w = i]$, где

$$[z] = \begin{cases} 1, & \text{если выражение z истинно,} \\ 0, & \text{иначе.} \end{cases}$$

Методы оценки качества решения

В области машинного обучения широко используется матрица неточностей (confusion matrix) [17], которая позволяет визуализировать результаты работы алгоритма. Эта матрица размера N на N, где N – это количество классов. Строки этой матрицы соответствуют истинным классам, а столбцы решениям классификатора. При классификации объекта из тестовой выборки мы увеличиваем число, стоящее на пересечении столбца класса, который вернул классификатор и строки класса к которому действительно относится данный объект. Название связано с тем, что такая матрица позволяет легко увидеть те классы, которые путает алгоритм.

Имея такую матрицу точность и полнота для каждого класса рассчитывается очень просто. Точность равняется отношению соответствующего диагонального элемента матрицы и суммы всего столбца класса. Полнота – отношению диагонального элемента матрицы и суммы всей строки класса:

$$\text{Precision}_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}},$$

$$\text{Recall}_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}}$$

Результирующая точность классификатора рассчитывается как арифметическое среднее его точности по всем классам. Аналогично вычисляется результирующая полнота.

В данной работе каждое значение confusion matrix делилось на сумму строки, чтобы сразу было понятно качество работы алгоритма.

Распознавание по данным от акселерометра

Для удаления неинформативных участков сигнал по каждой из осей x, y, z разбивался на окна. Считалась дисперсия сигнала D_x, D_y, D_z. Для каждого окна w_{ij}, где $i \in \{x, y, z\}$, j – это порядковый номер окна, считалась характеристика

$$\lambda_i = \begin{cases} 0, & \text{если дисперсия окна } w_{ij} \text{ в } \beta \text{ раз меньше, чем } D_i, \\ 1, & \text{иначе.} \end{cases}$$

Оставлялись только те окна, где $\lambda_x \lambda_y \lambda_z = 1$. Данный алгоритм позволяет убирать участки на которых колебания сигнала очень малы. Для подбора значения β вручную для нескольких человек были размечены участки на 2 класса: информативные и неинформативные. Экспериментальным путем было получено значение $\beta = 10$.

Для очистки сигнала от шума был выбран фильтр нижних частот с критической частотой равной $\alpha = 0.1$. На Рис. 4 показан пример применения этого фильтра. Синим цветом отмечены оригинальные значения. Красной линией соединены значения после фильтрации. Видно, что после фильтрации сигнал становится намного более гладким и становится возможным анализировать сигнал.

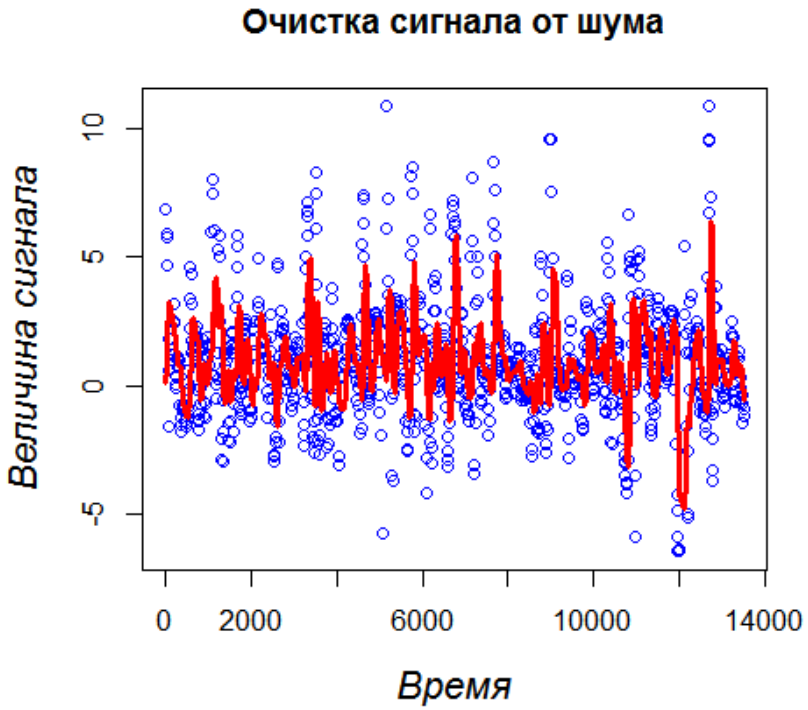


Рис. 4. Применение фильтра нижних частот

В ходе экспериментов были отобраны 8 лучших признаков:

— quantile0.1Z – квантиль порядка 0.1 для оси Z,
— meanY – среднее значение для оси Y,
— mobilityAH – mobility для горизонтальной проекции,
— dwa – Durbin-Watson autocorrelation,
— quantile0.25X – квантиль порядка 0.25 для оси X,
— complexityAV – complexity для вертикальной проекции,
— quantile0.1diff0.05Jerk – разность квантилей порядка 0.1 и 0.05 для фильтрации ударов,
— corrMagAv – корреляция между значениями общего ускорения и вертикальной проекции.

Таблица 4. Результаты работы алгоритма RF при использовании 8 признаков

actual/prediction	bag	hand	pocket
bag	0.647	0.160	0.192
hand	0.076	0.903	0.020
pocket	0.105	0.044	0.850

Таблица 5. Результаты работы алгоритма LDA при использовании 8 признаков

actual/prediction	bag	hand	pocket
bag	0.362	0.205	0.433
hand	0.032	0.927	0.041
pocket	0.158	0.076	0.766

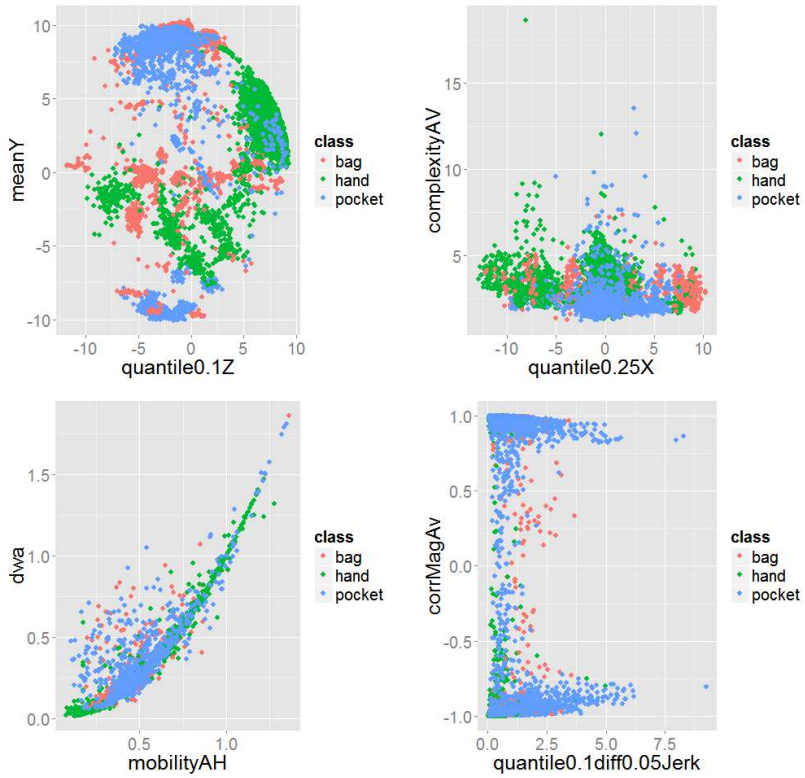


Рис. 5. Распределение объектов в пространствах 8 лучших признаков

Из графиков видим, что нахождение телефона в руке определяется практически без ошибок. Два оставшихся класса сильно перепутаны между собой. Это подтверждает матрица неточностей.

Распознавание по данным от акселерометра и гироскопа

Теперь попробуем использовать данные акселерометра и гироскопа вместе.

Таблица 6. Результаты работы алгоритма RF на данных акселерометра и гироскопа

actual/prediction	bag	hand	pocket
bag	0.7960	0.1040	0.1000
hand	0.0311	0.9654	0.0035
pocket	0.0189	0.0038	0.9773

Итоговый классификатор состоит из 18 признаков:

— Нормированное на длину окна количество пересечений квантиля уровня 0.25 для общего ускорения.
— Нормированное на длину окна количество пересечений квантиля уровня 0.5 для общего ускорения.
— Нормированное на длину окна количество пересечений квантиля уровня 0.5 для вертикальной проекции.
— Mobiliity для вертикальной проекции.
— Complexity для вертикальной проекции.
— Среднее значение мощности первой производной для горизонтальной проекции.

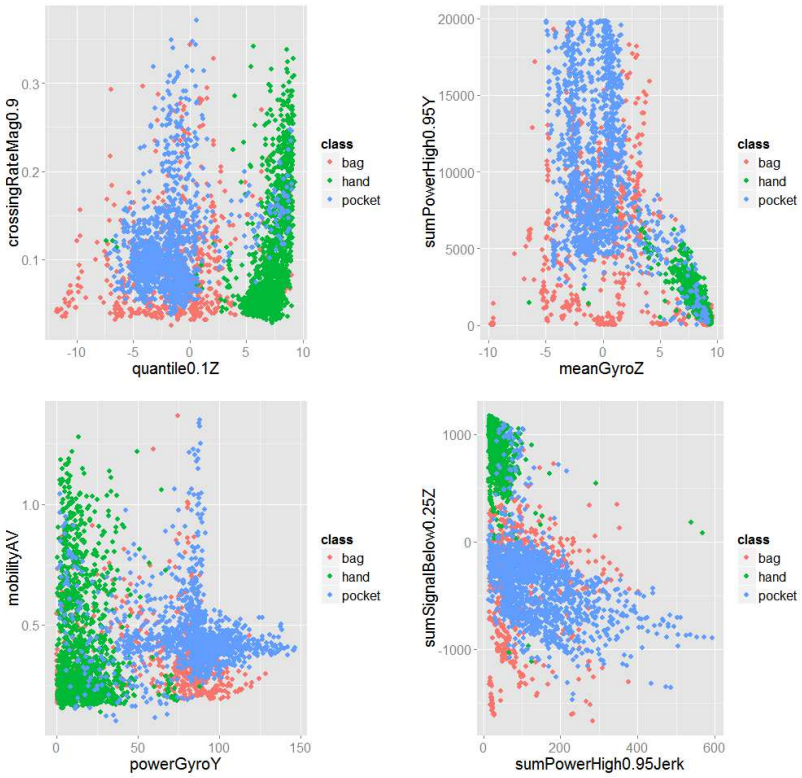


Рис. 6. Распределение объектов в пространствах лучших пар признаков от акселерометра и гироскопа

— Сумма значений сигнала, лежащих выше квантиля уровня 0.9 для оси Y.
— Мощность сигнала, лежащего выше квантиля уровня 0.9 для оси Y.
— Квантиль уровня 0.1 для оси Z.
— Сумма значений сигнала, лежащих ниже квантиля уровня 0.25 для оси Z.
— Сумма значений сигнала, лежащих выше квантиля уровня 0.95 для jerk-фильтра.
— Сумма значений сигнала, лежащих ниже квантиля уровня 0.1 для jerk-фильтра.
— Среднее значение показаний гироскопа для оси Y.
— Среднее значение показаний гироскопа для оси Z.
— Мощность сигнала гироскопа для оси X.
— Мощность сигнала гироскопа для оси Y.
— Durbin-Watson autocorrelation.
— Magnitude Mean Value.

Из графиков видно, что положение телефона в руке опять же определяется почти без ошибок по нескольким признакам. Зато теперь из таблицы неточностей замечаем, что и положение телефона в кармане определяется с 97% точностью.

Рассмотрим устойчивость решения. Для каждого значения количества деревьев будем производить 10 запусков и вычислять среднее качество классификации и стандартное отклонение. Под средним качеством подразумевается среднее значение диагональных элементов в матрице неточностей.

Видим, что при количестве деревьев $\geqslant 30$ качество классификации превышает 90% и стандартное отклонение между различными запусками не превосходит 0.005. Предполагая, что качество между различными запусками распределено нормально, находим, что среднее качество решения не менее 88 %.

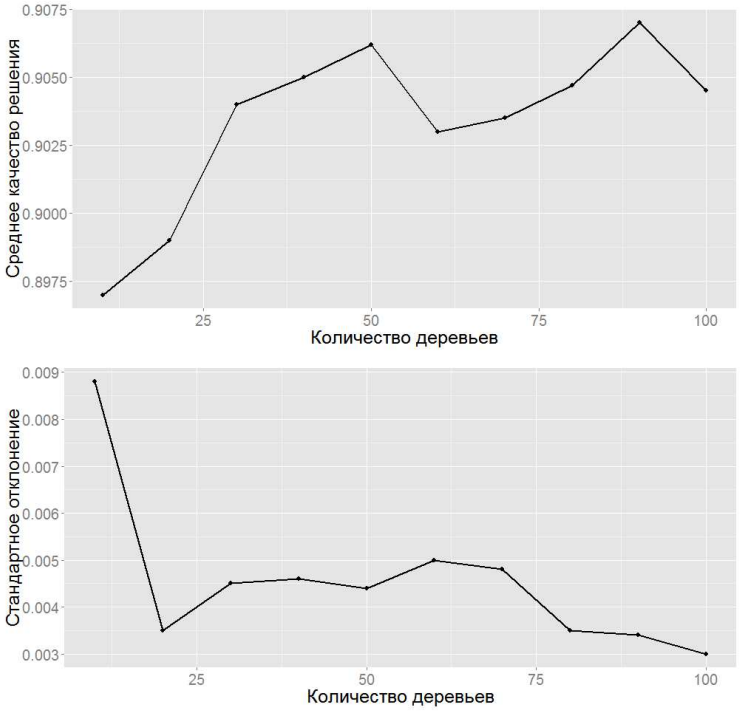


Рис. 7. Качество решение в зависимости от числа деревьев

Исходный код для генерации признаков находится в `https://github.com/MoRandi91/On-body-mobile-phone-localization`.

Заключение

В работе представлено экспериментальное исследование задачи определения местоположения телефона. Был произведен обзор алгоритмов вычисления признаков для сигналов трехосевого акселерометра, которые являются инвариантными к ориентации телефона. Описаны преимущества совместного использования данных акселерометра и гироскопа. Построен алгоритм, который более чем с 97% точностью определяет положение телефона в кармане и в руке. Общее качество построенного алгоритма порядка 88%. Используя полученные результаты, можно модифицировать методы решения задачи и улучшить их производительность. В работе не затронут анализ влияния параметров алгоритмов предварительной обработки и вычисления признаков, таких как частота дискретизации, размер окна, доля перекрытия окон, на качество классификации, что может послужить темой для дальнейших работ.

Автор выражает глубокую признательность компании Nokia и лично Сафонову Илье Владимировичу за предоставленные для экспериментов данные, консультации при проведении экспериментов и ценные замечания, способствовавшие улучшению работы. Также автор выражает признательность своему научному руководителю Александру Геннадьевичу Дьяконову, который контролировал весь процесс работы с задачей.

Литература

[1] *Bao L., Intille S. S.* Activity Recognition from User-Annotated Acceleration Data // *In Proc. of the 2nd International Conference on Pervasive Computing (Pervasive 2004).* Vienna, 2004. P. 1–17.

[2] *Lee S.-W., Mase K.* Recognition of walking behaviors for pedestrian navigation // *In Proc. of 2001 IEEE Conference on Control Applications (CCA 2001).* Los Alamitos, 2001. P. 1152–1155.

[3] *Sumida M., Mizumoto T., Yasumoto K.* Smartphone-based heart rate prediction for walking support application // *IEICE Technical Report*, 2013.

[4] *Fujinami K., Kouchi S.* Recognizing a Mobile Phone's Storing Position as a Context of a Device and a User // *Mobile and Ubiquitous Systems: Computing, Networking, and Services., Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2013. Vol. 120. P. 76–88.

[5] *Vahdatpour A., Amini N., Sarrafzadeh M.* On-body Device Localization for Health and Medical Monitoring Applications // *In Proc. of the 2011 IEEE International Conference on Pervasive Computing and Communications (PERCOM '11).* Seattle, 2011. P. 37–44.

[6] *Dixon B. C.* Zero-gravity maneuver instruments and instrumentation. Defense Technical Information Center, 1966. 60 p.

[7] *Crossbow, Inc.* Rate Gyro Application Note: Theory of Operation of Angular Rate Sensors. Available at: `www.moog-crossbow.com/Literature/Application_Notes_Papers/Theory_of_Operation_of_Rate_Sensors.pdf`

[8] *Brajdic A., Harle R.* Walk Detection and Step Counting on Unconstrained Smartphones // *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing(UbiComp'13).* Zurich, 2013. P. 225–234.

[9] *Cobb J. E.* An accelerometer based gestural capture system for performer based music composition. MSc by Research in Music Technology. University of York, 2011.

[10] *Stoichkov R.* Android Smartphone Application for Driving Style Recognition. Lehrstuhl fur Medientechnik, Technische Universitat Munchen, 2013.

[11] *M. Khan, S. I. Ahamed, M. Rahman, R. O. Smith* A Feature Extraction Method for Realtime Human Activity Recognition on Cell Phones // *Proc. of 3rd International Symposium on Quality of Life Technology (isQoLT 2011).* Toronto, 2011.

[12] *Batist P., Silvestre C., Oliveira P., Cardeira B.* Accelerometer Calibration and Dynamic Bias and Gravity Estimation: Analysis, Design, and Experimental Evaluation // *IEEE Trans. Contr. Sys. Techn*, 2011. Vol. 9, No. 5. P. 1128–1137.

[13] *Hjorth B.* EEG analysis based contributions on time domain properties // *Eleetroeneeph. clin. Neurophysiol*, 1970. Vol. 29. P. 306–310.

[14] *Durbin J., Watson G. S.* Testing for Serial Correlation in Least Squares Regression // *Biometrika*, 1950. Vol. 34, No. 3–4. P. 409–428.

[15] *Ustev Y. E., Ersoy C., Incel O. D.* User, Device and Orientation Independent Human Activity Recognition on Mobile Phones: Challenges and a Proposal // *Proc. of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp '13 Adjunct).* New York, 2013. P. 1427–1436.

[16] *Ravi N., Dandekar N., Mysore P., Littman M. L.* Activity Recognition from Accelerometer Data // *Proc. of the 17th conference on Innovative applications of artificial intelligence (IAAI'05).* Pittsburgh, 2005. P. 1541–1546.

[17] *Stehman S. V.* Selecting and interpreting measures of thematic classification accuracy // *Remote Sensing of Environment*, 1997. Vol. 62, No. 1. P. 77–89.

Оценки, минимизирующие возможность потерь, и минимаксные оценки: сравнительный анализ*

А. И. Чуличков, Б. Юань
achulichkov@gmail.com
Физический Факультет МГУ им. М.В.Ломоносова, Москва, Ленинские горы, Дом 1, строение 2

Поставлена и решена задача оценивания значений функции в заданных точках области ее определения по результатам измерений конечного набора ее функционалов, выполненных с погрешностью. Показано, что с конечной погрешностью может быть оценена только конечномерная составляющая искомой функции, предложена точная конечномерная модель, позволяющая построить искомые оценки. Обсуждаются два метода оценивания. Первый метод минимизирует максимально возможную погрешность оценивания каждого значения функции в заданной точке. Считается, что погрешность измерения каждого линейного функционала с одной и той же возможностью принимают любое значение внутри заданного интервала. Для каждого оцениваемого значения функции построен интервал, которому может принадлежать это значение. Минимаксной оценкой является середина этого интервала, а погрешностью оценки – половина его длины. Концы каждого интервала определяются решениями задач линейного программирования. Второй метод оценивания основан на теоретико-возможностной модели измерений, в которой считается, что большие значения погрешности измерения каждого функционала менее возможны, чем малые. Критерием оценивания является возможность потерь. Метод оценивания минимизирует этот критерий и сводится к решению задачи линейного программирования. Оценки минимальной возможности потерь сравниваются с оценками, минимизирующими максимальную погрешность каждого значения функции. Обсуждаются различия минимаксных оценок и оценок минимальной возможности потерь. Приведен пример оценивания параметров реального спектрометрического эксперимента.

Ключевые слова: минимаксные оценки; теория возможностей; оценки минимальной возможности потерь; линейное программирование; анализ экспериментальных данных.

Estimation of minimum possibility of losses and minimax estimation: a comparative analysis*

A. Chulichkov, B. Yuan
Faculty of Physics M.V. Lomonosov Moscow State University, Moscow, Russia

The estimation of function values at specified points in its domain of definition based on the measurement results of the finite set of functionals is posed and solved. The measurements are distorted by a finite error. It is shown that with the finite error can be estimated only finite-dimensional component of unknown function. Exact finite-dimensional model, underlying the construction of required assessments is proposed. Two methods for estimation are discussed. The first method minimizes the maximal error of the estimation of each value of the function at a given point. It is believed that the measurement error of each linear functional may take any value within a given interval. For each of the estimated value of a function the interval that contains this value was constructed. The minimax estimate is the midpoint of this interval, and the error is the half of its length. The ends of each interval are determined

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-07-00409

as solutions of linear programming problems. The second method of estimation is based on theoretical-possibility measurement models. It is believed that large values of the measurement error of each functional less possible than small. The estimation criterion is the possibility of losses. Estimation method minimizes this criterion and is reduced to the solution of a linear programming problem. The estimates of the minimum possible losses and the estimates that minimize the maximum error of each value of function is compared. Differences between the minimax estimates and estimates of the minimum possible losses are discussed. An example of estimation of specter based on the data of real spectrometric experiment is given.

Keywords: minimax estimation; the theory of possibilities; estimates of the minimum possible losses; linear programming; analysis of experimental data.

Введение

В современных экспериментальных исследованиях часто возникает необходимость восстанавливать характеристики исследуемых объектов по данным измерений, причем изучаемые характеристики лишь косвенно связаны с результатом измерительных экспериментов. Для таких исследований типичной является схема

$$\xi = Ag + \nu, \tag{1}$$

интерпретируемая следующим образом: результат измерения ξ есть искаженный шумом ν выходной сигнал Ag измерительного прибора A, на вход которого подан сигнал g от измеряемого объекта. Задача интерпретации измерения (1) состоит в том, чтобы извлечь из ξ наиболее точную версию входного сигнала g или вычислить значение заданной функции Ug. Исходными данными для решения такой задачи является математическая модель схемы измерений (1) и данные измерений ξ.

Примером таких экспериментов является измерение спектра электромагнитного излучения с помощью спектрометра [1]. В этом случае входным сигналом g является спектр электромагнитного излучения $g(\cdot)$, выходной сигнал $q = Ag$ спектрометра формируется согласно соотношению

$$q(\lambda) = \int_0^\infty a(\lambda, \lambda')g(\lambda')d\lambda', \quad \lambda \in [0, \infty). \tag{2}$$

Здесь $a(\cdot, \cdot)$ — аппаратная функция спектрометра, ее смысл состоит в том, что при подаче на вход спектрометра монохроматического спектра единичной интенсивности с длиной волны λ' на выходе спектрометра получим спектр $a(\lambda, \lambda')$, $\lambda \in [0, \infty)$.

Одним из широко распространенных подходов к решению задачи интерпретации измерений (1) состоит в решении интегрального уравнения Фредгольма 1 рода (2) на основании известной аппаратной функции $a(\cdot, \cdot)$ и данных измерений функции $q(\cdot)$, выполненных с погрешностью ν; погрешность ν при этом считается либо ограниченной по норме [2], либо обладающей известными стохастическими свойствами [3, 4]. Однако, как показано в работах [2, 5, 6], такая задача может оказаться некорректно поставленной по Адамару, в частности, может оказаться неразрешимой, либо иметь неединственное решение, либо ее решение (псевдорешение [5], если уравнение $Ag = \xi$ неразрешимо) обладает неустойчивостью по отношению к возмущению функции $q(\cdot)$ или математической модели измерений. Для решения некорректно поставленных задач были предложены методы регуляризации,

основная идея которых состоит в наложении дополнительных ограничений на класс решений, в результате решение регуляризованной задачи оказывается единственным, и стремится к точному при стремлении погрешности измерения ν к нулю [7, 8, 9, 10]. Несмотря на значительные успехи в создании методов решения таких задач, интерес к ним не угасает и в настоящее время [11, 12].

Поиск регуляризованного решения часто производится минимизацией суммы функционала невязки $\|Ag - \xi\|$ и сглаживающего функционала $\Omega(g)$:

$$g_\alpha = \arg\min\{\|Ag - \xi\| + \alpha\Omega(g)\}, \quad \alpha \geqslant 0, \tag{3}$$

так, чтобы при согласованном стремлении $\alpha \to 0$ и $\nu \to 0$ решение g_α стремится к точному решению (или к псевдорешению).

Такой подход к решению задачи интерпретации измерений встречает ряд принципиальных трудностей. Во-первых, как правило, экспериментатора интересует наиболее *точная оценка* сигнала g (или сигнала $u = Ug$, где U — заданное преобразование). В задаче (3) минимизируемый функционал не имеет отношения к точности решения. Во-вторых, в реальных ситуациях погрешность не стремится к нулю, и возникают вопросы с выбором параметра $\alpha > 0$.

Другой подход к решению задач интерпретации измерений дает теория измерительно-вычислительных систем (ИВС), развиваемая в школе проф. Ю.П.Пытьева [13]. В теории ИВС эти задачи ставятся как задачи поиска такого преобразования R измерения ξ, результат которого наиболее близок к $u = Ug$. При этом сигнал $u = Ug$ интерпретируется как выходной сигнал Ug «идеального» прибора U, на вход которого подан сигнал g. Конкретная постановка задач оценивания сигнала u зависит от математической модели измерения (1) и критериев точности оценки. Например, если задать модель измерения (1), указав множества, которым принадлежат сигналы g и ν, $g \in G$, $\nu \in N$, то оценкой $R\xi = \widehat{u}$ сигнала $u = Ug$ является

$$\widehat{u} = \arg\inf_u \sup_g \{\|u - Ug\| \,|\, \xi = Ag + \nu, \nu \in N, g \in G\}. \tag{4}$$

Оценка $\widehat{u}$ называется минимаксной, она минимизирует максимально возможную погрешность оценки u [13, 14].

Однако минимаксные оценки в силу постановки задачи (4) минимизируют ошибку в самых неблагоприятных ситуациях, которые, на взгляд исследователя, могут показаться не вполне реальными. Поясним это на простейшем примере.

Пусть в (1) результатом измерения является вектор ξ линейного пространства размерности, равной двум, с координатами (0.5,3), линейный оператор A задан матрицей $A = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$, на координаты векторов g и ν наложены ограничения: $|\nu_i| \leqslant 1$, $g \in G = \{(g_1, g_2) : 0 \leqslant g_i \leqslant 1, i = 1, 2\}$. На рисунке 1, слева, показана область возможных значений вектора g в виде единичного квадрата. Образ этой области при отображении A показан на рисунке 1, справа, в виде параллелограмма. Точкой обозначен вектор ξ с координатами $(0.5, 3)$. Координаты вектора Ag отличаются от координат ξ не более, чем на единицу, с другой стороны, принадлежат образу множества G при отображении A — т.е. области на плоскости в виде треугольника ABC. Полный прообраз этой области показан на рисунке 1, слева, в виде треугольника $A'B'C'$. Таким образом, результат измерения $\xi = (0.5, 3)$ в рамках принятой математической модели измерения, если и только если первая координата вектора g изменяется от 0 до 1, а вторая — от 0.5 до 1. Минимаксная

оценка первой координаты вектора g равна $\hat{g}_1 = 0.5$ с погрешностью 0.5, а вторая $\hat{g}_2 = 0.75$ с погрешностью 0.25. Однако образ $A\hat{g} = (0.5, 2)$ отличается от результата измерения $\xi = (0.5, 3)$ на вектор, одна из координат которого равна 1. Это означает, что получить результат ξ в измерении (1) при $g = \hat{g}$ можно только при максимальной измерительной погрешности второй координаты вектора Ag.

Реализация максимальной измерительной погрешности может показаться исследователю весьма неправдоподобной, поскольку измерительные эксперименты стараются организовать так, чтобы погрешности измерения были как можно меньше.

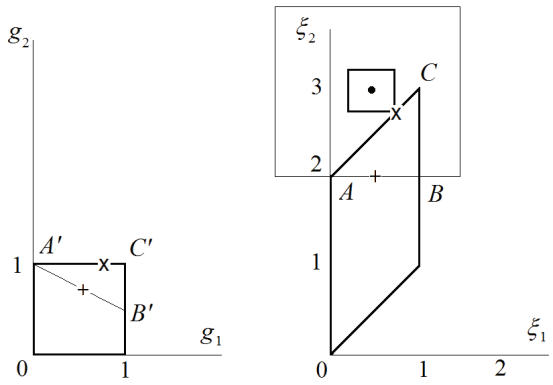


Рис. 1. Минимаксное оценивание вектора g в (1) и оценка, учитывающая, что малые значения погрешности измерений более возможны, чем большие

Учесть последнее замечание можно, отказавшись от предположения о том, что нет никаких предпочтений в значениях погрешности измерений, и указав, какие из их значений более возможны (правдоподобны), а какие — менее. В частности, если считать, что малые значения координат погрешности более возможны, чем большие, то точка Ag из треугольника ABC, координаты которой наиболее близки к координатам ξ, соответствует измерению с минимальными координатами измерительной погрешности. В этом смысле эта точка Ag наиболее правдоподобна из всех точек треугольника ABC. На рисунке 1, справа, это точка касания квадрата с центром в точке ξ с треугольником ABC, ее прообраз показан на рисунке 1, слева, знаком "x".

Математический формализм, позволяющий при оценивании параметров учесть предпочтения в значениях параметров математической модели измерения, дается в теории возможностей [15].

Впервые теория возможностей, основанная на теории нечетких множеств [16], была предложена Л.Заде в работе [17]. В отличие от «четкого» множества A, задаваемого индикатором $\chi_A(\cdot)$: $\chi_A(x) = 1$ для $x \in A$ и $\chi_A(x) = 0$ для $x \notin A$, нечеткое множество задается функцией принадлежности $\mu_A(\cdot)$, принимающей значение на отрезке $[0, 1]$, при этом если $\mu_A(x) = 0$, то элемент x не может принадлежать нечеткому множеству A, и чем больше $\mu_A(x)$, тем более возможным является утверждение, что $x \in A$; равенство $\mu_A(x) = 1$, означает, что включение $x \in A$ вполне возможно. Функция принадлежности интерпретируется как функция распределения возможностей включения x в A. Далее математическая теория возможностей развивалась в работах Д. Дюбуа, Г.Прадта [18, 19], Р. Ягера [20], Ю.П.Пытьева [15] и др.

В работах Ю.П.Пытьева предлагается трактовать возможность события как относительную оценку истинности данного события, его предпочтительности в сравнении с любым другим. Значение возможности события используется лишь для того, чтобы сравнить ее с возможностью любого другого события, указав, какое из двух более возможно или заключив, что они равновозможны. Это существенно расширяет возможности моделирования нечеткости, так как все распределения, значения которых связаны строго монотонно возрастающим преобразованием, оказываются эквивалентными. В отличие от вероятности, возможность не имеет событийно-частотной интерпретации, которая связывает её с экспериментом. Тем не менее теория возможностей позволяет математически моделировать реальность на основе опытных фактов, знаний, гипотез, суждений исследователей.

Математической моделью нечеткости в [15] является пространство с возможностью $(\Omega, \mathrm{P}(\Omega), P)$, где Ω есть пространство элементарных событий, $\mathrm{P}(\Omega)$ есть алгебра всех подмножеств Ω, и $P(\cdot) : \mathrm{P}(\Omega) \to [0,1]$ — возможностная мера. Возможность $P(\cdot)$ определена как мера на $\mathrm{P}(\Omega)$, и принимает значение на множестве $[0,1]$, на котором определены бинарные операции «сложения» $+$ и «умножения» $\bullet$ согласно равенствам $a+b = \max\{a,b\}$, $a \bullet b = \min\{a,b\}$, $(a,b) \in [0,1] \times [0,1]$. В работе [15] показано, что возможность любого события $A \in \mathrm{P}(\Omega)$ может быть представлена в виде

$$P(A) = \sup_x \min\{p(x), \chi_A(x)\}, \tag{5}$$

где $\chi(\cdot)$ — индикатор A, $p(\cdot)$ — распределение возможностей, определенное как возможность одноточечного события $p(x) = P(\{x\})$, $x \in \Omega$. Формула (5), как и многие другие факты теории возможностей, получены из аналогичных формул теории вероятностей заменой операции интегрирования на sup, суммы на max, умножения — на min. Так, например, события A и B независимы в теоретико-возможностном смысле, если $P(AB) = \min\{P(A), P(B)\}$; заметим, что это отличает рассматриваемый вариант теории возможностей от теории Л.Заде, в которой последнее равенство постулируется для всех событий A, B.

В теории возможностей аналогом случайных элементов некоторого линейного нормированного пространства $\mathcal{R}$ являются нечеткие элементы, определенные как любая функция $\varphi(\cdot)$, заданная на Ω и принимающая значение в $\mathcal{R}$. Нечеткий элемент φ характеризуется распределением возможностей своих значений, заданным функцией $\pi^{\varphi}(\cdot)$, определенной на $\mathcal{R}$ и принимающей значения в отрезке $[0,1]$, значение $\pi^{\varphi}(y) = P(\{\varphi = y\})$ по определению есть возможность того, что нечеткий элемент φ примет значение y, $y \in \mathcal{R}$. На функцию $\pi^{\varphi}(\cdot)$ накладывается условие нормировки: $\max\limits_{y \in \mathcal{R}}\{\pi^{\varphi}(y)\} = 1$.

Для того, чтобы формализовать утверждение «малые погрешности измерений более возможны, чем большие», достаточно считать погрешность $\nu \in \mathcal{R}_1$ в (1) нечетким элементом линейного нормированного пространства $\mathcal{R}_1$ и задать распределение возможностей его значений $\nu = z$ как монотонно убывающую функцию нормы $z \in \mathcal{R}_1$: $\pi^{\nu}(z) = \pi_0(\|z\|)$, $z \in \mathcal{R}_1$. Априорное знание $g \in G \subset \mathcal{R}_2$ в (4) о возможных значениях элемента $g \in \mathcal{R}_2$ может быть задано распределением возможности $\pi^{g}(y) = \chi_G(y)$, $y \in \mathcal{R}_2$. Если g и ν независимы, то совместное распределение возможностей пары $(g, \nu) \in \mathcal{R}_2 \times \mathcal{R}_1$ дается равенством $\pi^{g,\nu}(y,z) = \min\{\pi^{g}(y), \pi^{\nu}(z)\} = \min\{\chi_G(y), \pi_0(\|z\|)\}$. Если выполнено (1), то совместное

распределение тройки $(\xi, g, \nu) \in \mathcal{R}_2 \times \mathcal{R}_2 \times \mathcal{R}_1$ дается равенством

$$\pi^{\xi,g,\nu}(x,y,z) = \begin{cases} \min\{\pi^g(y), \pi^\nu(z)\}, & \text{если } x = Ay + z, \\ 0, & \text{если } x \neq Ay + z, \end{cases} =$$
$$= \begin{cases} \min\{\chi_G(y), \pi_0(\|x - Ay\|)\}, & \text{если } x = Ay + z, \\ 0, & \text{если } x \neq Ay + z. \end{cases}$$

Распределение возможности пары нечетких элементов ξ, g дается "интегрированием":

$$\pi^{\xi,g}(x,y) = \sup_z \pi^{\xi,g,\nu}(x,y,z) = \min\{\chi_G(y), \pi_0(\|x - Ay\|)\}. \tag{6}$$

В общем случае в [15] решение о значении оцениваемой величины g определяется как нечеткий элемент δ с распределением переходной возможности $\pi^{\delta|\xi}(\cdot|\cdot)$, так, что при заданном значении $\xi = x$ значение $\pi^{\delta|\xi}(d|x)$ есть возможность считать d оценкой для g. Однако на практике в ряде случаев можно использовать "четкое" правило оценивания, задав функцию $d(\cdot)$, которая каждому значению $\xi = x$ ставит в соответствие оценку $d(x)$ нечеткого элемента g.

В [15] с каждым значением d оценки g предлагается связать значение возможности потерь $l(d, g)$, возникающих при использовании оценки d вместо истинного значения g. Тогда возможность потерь при решающем правиле $d(\cdot)$ равна $PL(d(\cdot)) = \sup_{x,g} \min\{l(d(x), g), \pi^{\xi,g}(x,y)\}$. Оптимальная оценка, минимизирующая возможность потерь, есть решение задачи на минимум

$$PL(d_*(\cdot)) = \arg\min_{d(\cdot)} PL(d(\cdot)). \tag{7}$$

В [15] показано, что оптимальную оценку можно получить, решая задачу (7) при каждом значении $\xi = x$.

В настоящей статье на примере решения задачи интерпретации данных спектрометрического эксперимента сравниваются два подхода, один из них основан на минимаксном оценивании параметров входного спектра, другой — на минимизации возможности потерь. В работе решение задачи интерпретации данных эксперимента (1) сводится к конечномерным задачам линейного программирования, при этом вместо замены операции интегрирования приближенной суммой, широко используемой при решении такого типа задач, предлагается точный метод дискретизации разложением по системе функций, полной в подпространстве, ортогональном нуль-пространству линейного интегрального оператора в (2). Показано, что в задачах интерпретации данных спектрометрии минимаксный подход не дает удовлетворительной оценки входного спектра, однако предположение о том, что большие значения измерительной погрешности менее возможны, чем малые, существенно улучшают оценку.

Математическая модель измерений

Пусть в (2) функции $g(\cdot)$, $a(\lambda, \cdot)$ являются элементами евклидова пространства $\mathcal{L}^2([0,\infty))$ при каждом $\lambda \in [0,\infty)$. Однако на практике в экспериментальных исследованиях измеряется не функция $q(\lambda)$, $\lambda \in [0,\infty)$, а конечный набор чисел. Если функция $a(\cdot,\cdot)$ задана своим представителем, являющимся непрерывной функцией на $[0,\infty)\times[0,\infty)$, то эти числа рассматриваются как значения функции $q(\cdot)$ в точках $\lambda_1, \ldots, \lambda_n$. Результаты измерения значений $q(\lambda_i)$ сопровождаются погрешностью (шумом) ν_i, $i = 1, \ldots, n$. В

частности, в рассматриваемом примере результатом измерения спектра являются числа $\xi_1, \ldots, \xi_n$, интерпретируемые как искаженные шумом значения спектра на выходе спектрометра на длинах волн $\lambda_1, \ldots, \lambda_n$:

$$\xi_i = q(\lambda_i) + \nu_i = \int_0^\infty a(\lambda_i, \lambda')g(\lambda')d\lambda' + \nu_i, \quad \lambda \in [0, \infty). \tag{8}$$

Заметив, что $\int_0^\infty a(\lambda_i, \lambda')g(\lambda')d\lambda' = (a_i, g)$, где $a_i \in \mathcal{L}^2([0, \infty))$ представлен непрерывной функцией $a(\lambda_i, \cdot)$, $i = 1, \ldots, n$, а $(\cdot, \cdot)$ — скалярное произведение в $\mathcal{L}^2([0, \infty))$, запишем (8) в виде

$$\xi_i = (a_i, g) + \nu_i, \quad i = 1, \ldots, n, \tag{9}$$

элементы $a_1, \ldots, a_n$ считаем линейно независимыми. Погрешности измерений ν_i принадлежат заданным интервалам:

$$|\nu_i| \leqslant \varepsilon_i, \quad i = 1, \ldots, n. \tag{10}$$

Как видно из (9), оценке с конечной погрешностью поддается лишь проекция $g \in \mathcal{L}^2([0, \infty))$ на линейную оболочку S_A элементов $a_1, \ldots, a_n \in \mathcal{L}^2([0, \infty))$, совпадающую с ортогональным дополнением к нуль-пространству интегрального оператора в (8). Поскольку $a_i = a(\lambda_i, \cdot)$, $i = 1, \ldots, n$, непрерывны, то элемент $Pg \in S_A$ можно задать его представителем $Pg(\cdot) = \sum_{i=1}^n \varphi_i a(\lambda_i, \cdot)$, для которого определены его значения в любой точке $\lambda' \in [0, \infty)$. Здесь $\varphi_1, \ldots, \varphi_n$, — коэффициенты разложения проекции $Pg \in \mathcal{L}^2([0, \infty))$ по системе линейно независимых элементов $a(\lambda_i, \cdot) \in \mathcal{L}^2([0, \infty))$, $i = 1, \ldots, n$.

Представив в (9) g в виде $g = Pg + (I - P)g$ и учтя, что $(I - P)g$ ортогонально всем a_i, $i = 1, \ldots, n$, получим $\xi_i = \sum_{j=1}^n (a_i, a_j)\varphi_j + \nu_i$. Итак, результат измерения определяется значением конечного числа параметров $\varphi_1, \ldots, \varphi_n$. Теперь ξ_i в (9) удобно представить как искаженный шумом ν_i результат измерения скалярного произведения двух n-мерных векторов $\vec{\varphi}$ и $\vec{a}_i$, заданных своими координатами: $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$, $\vec{a}_i = ((a_i, a_1), \ldots, (a_i, a_n))$, $i = 1, \ldots, n$:

$$\xi_i = (\vec{a}_i, \vec{\varphi})_n + \vec{\nu}, \tag{11}$$

здесь $(\cdot, \cdot)_n$ — скалярное произведение в R^n.

Пусть требуется оценить значение проекции $Pg(\cdot)$ в точках $\lambda'_1, \ldots, \lambda'_N \in [0, \infty)$. Эти значения будем рассматривать как координаты u_j вектора $\vec{u} \in \mathcal{R}^N$. Тогда

$$u_j = Pg(\lambda_j) = \sum_{i=1}^n \varphi_i a(\lambda_i, \lambda_j), \quad j = 1, \ldots, N, \tag{12}$$

здесь $\varphi_i, i = 1, \ldots, n,$ — коэффициенты разложения Pg по линейно независимым элементам a_i, $i = 1, \ldots, n$. Введя векторы $\vec{w}_j \in \mathcal{R}^n$, $i = j, \ldots, N$, заданные своими координатами $\vec{w}_j = (a(\lambda_j, \lambda_1), \ldots, a(\lambda_j, \lambda_n))$, $j = 1, \ldots, N$, запишем координаты вектора $\vec{u} \in \mathcal{R}^N$ как скалярные произведения

$$u_j = (\vec{w}_j, \vec{\varphi})_n, \quad j = 1, \ldots, N. \tag{13}$$

Оценка значений $Pg(\lambda_j)$, $j = 1, \ldots, N$, проекции элемента $g \in \mathcal{L}^2([0,\infty))$ на S_A в точках $\lambda'_1, \ldots, \lambda'_N \in [0,\infty)$ теперь получается оцениванием конечного числа координат (13) вектора $\vec{u} \in \mathcal{R}^N$ по результату (11) измерений n линейных функционалов конечномерного вектора $\vec{\varphi} \in \mathcal{R}^n$. Априорные ограничения на искомую оценку спектра (например, неотрицательность координат вектора $\vec{u}$) могут быть заданы в виде линейных неравенств

$$l_j \leqslant u_j \leqslant r_j, \ -\infty \leqslant l_j < r_j \leqslant \infty, \ \ j = 1, \ldots, N. \tag{14}$$

В частности, если известно, что все координаты $u_1, \ldots, u_N$ неотрицательны, то $l_j = 0$, $r_j = +\infty$, $j = 1, \ldots, N$.

Минимаксная оценка значения координат вектора $\vec{u}$

Рассмотрим задачу, в которой оценка каждой координаты u_j вектора $\vec{u}$ определяется из принципа минимизации максимальной погрешности оценки.

Согласно (13) $u_j = (\vec{w}_j, \vec{\varphi})_n$, а $\vec{\varphi} \in \mathcal{R}^n$ в силу (9) и (10) удовлетворяет системе линейных неравенств

$$|(\vec{a}_i, \vec{\varphi})_n - \xi_i| \leqslant \varepsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

и системе линейных неравенств априорных ограничений (14)

$$l_k \leqslant (\vec{w}_k, \vec{\varphi})_n \leqslant r_k, \quad k = 1, \ldots, N, \tag{16}$$

Решение системы неравенств (15) и (16) есть выпуклое замкнутое множество, и множество значений проекций всех элементов этого множества на любой вектор $\vec{w}_j \in \mathcal{R}^n$ есть отрезок. Поэтому искомая координата $u_j = (\vec{w}_j, \vec{\varphi})_n$ принадлежит интервалу $u_j \in [u_{j,\min}, u_{j,\max}]$, левый и правый края которого определяются как минимальное и максимальное значение скалярного произведения $(\vec{w}_j, \vec{\varphi})_n$ при ограничениях, заданных системой линейных неравенств (15)-(16), т.е. решениями соответствующих задач линейного программирования.

Определив границы соответствующих интервалов, минимаксную оценку $\hat{u}_j$ каждой координаты u_j, $j = 1, \ldots, N$, вектора $\vec{u}$ найдем из задачи на минимакс

$$\hat{u}_j = \arg\min_{u'_j} \max_{u_{j,\min} \leqslant u_j \leqslant u_{j,\max}} |u_j - u'_j|.$$

Ее решение, очевидно, есть середина интервала $[u_{j,\min}, u_{j,\max}]$, а погрешность — половина его длины: $\hat{u}_j = (u_{j,\min} + u_{j,\max})/2$, $h_j = (u_{j,\max} - u_{j,\min})/2$, $j = 1, \ldots, N$.

Оценки максимальной возможности

Как показано в предыдущих разделах настоящей статьи, задача интерпретации измерений (8) сводится к задаче оценивания значений проекции элемента $g \in \mathcal{L}^2([0,\infty))$ на линейную оболочку S_A линейно независимых элементов $a_1, \ldots, a_n$, заданных непрерывными функциями; ищется оценка значений проекции в точках $\lambda_1, \ldots, \lambda_N \in [0,\infty)$. Эти значения, рассматриваемые как координаты вектора $\vec{u} \in \mathcal{R}^N$, связаны линейным преобразованием $U \in (\mathcal{R}^n \to \mathcal{R}^n)$ с вектором $\vec{\varphi} \in \mathcal{R}^n$. Матрица этого линейного оператора задана в (13): $U_{j,i} = a(\lambda_j, \lambda_i)$, $i = 1, \ldots, n$, $j = 1, \ldots, N$. Вектор $\vec{\varphi}$ связан с вектором

$\vec{\xi} \in \mathcal{R}^n$ (с результатом измерения) соотношением (11). Таким образом, если задать векторы $\vec{\xi}, \vec{\nu} \in \mathcal{R}^n$ координатами $\vec{\xi} = (\xi_1, \ldots, \xi_n)$, $\vec{\nu} = (\nu_1, \ldots, \nu_n)$ и определить линейный оператор $A \in (\mathcal{R}^{\rightarrow}\mathcal{R}^n)$ его матрицей $A_{ik} = (a_i, a_k)$, $i, k = 1, \ldots, n$, то получим соотношения

$$\vec{\xi} = A\vec{\varphi} + \vec{\nu}, \quad \vec{u} = U\vec{\varphi}. \tag{17}$$

Задача состоит в том, чтобы, зная математическую модель измерения (17) и его результат $\vec{\xi}$, найти оценку вектора $\vec{u}$.

В соответствии с описанной во введении методикой оценивания, минимизирующей возможность потерь, зададим, во-первых, нечеткую модель погрешности измерения $\vec{\nu}$. Считая его координаты ν_i, $i = 1, \ldots, n$, независимыми нечеткими величинами, зададим распределение возможностей $\pi^{\nu_i}(\cdot)$ для каждой из них соотношением

$$\pi^{\nu_i}(z) = \begin{cases} \pi_0\left(\frac{|z|}{\varepsilon_i}\right), & |z| \leqslant \varepsilon_i, \\ 0, & |z| > \varepsilon_i, \end{cases} \tag{18}$$

где $\pi_0(\cdot)$ монотонно убывает на интервале $[0, 1]$. Это есть формальное выражение априорной уверенности экспериментатора в том, что большие погрешности менее возможны, чем малые, а также в том, что ошибки ν_i, по модулю превосходящие ε_i, невозможны. Тогда совместное распределение возможностей нечеткого вектора погрешности $\vec{\nu} = (\nu_1, \ldots, \nu_n)$ равно $\pi^{\nu}(\vec{z}) = \min\{\pi^{\nu_1}(z_1), \ldots, \pi^{\nu_n}(z_n)\} = \pi_0(\max\limits_{i=1,\ldots,n} \frac{|z_i|}{\varepsilon_i})$. Заметим, что $\max\limits_{i=1,\ldots,n} \frac{|z_i|}{\varepsilon_i}$ как функция вектора $\vec{z} = (z_1, \ldots, z_n)$ обладает всеми свойствами нормы $\vec{z}$, и будем в дальнейшем использовать для этой функции обозначение $\|\vec{z}\|_{\varepsilon}$. Таким образом, $\pi^{\nu}(\vec{z}) = \pi_0(\|\vec{z}\|_{\varepsilon})$.

Во-вторых, зададим априорное распределение возможностей нечеткого вектора $\vec{\varphi}$ как индикатор подмножества векторов $\mathcal{R}^n$, координаты которых удовлетворяют неравенствам (16):

$$\pi^{\varphi}(\vec{y}) = \begin{cases} 1, & \text{если } l_k \leqslant (\vec{w}_k, \vec{y})_n \leqslant r_k, \quad k = 1, \ldots, N, \\ 0, & \text{в противном случае.} \end{cases}$$

Это есть формальное выражение априорной уверенности исследователя в том, что значения искомой проекции элемента g на S_A принадлежат "четкому" множеству, заданному системой линейных неравенств (16).

Эти две модели с учетом (17) и в предположении независимости нечеткой погрешности измерений $\vec{\nu}$ и $\vec{\varphi}$ позволяют записать совместное распределение возможности пары нечетких векторов $\vec{\xi}$ и $\vec{\varphi}$ в виде $\pi^{\xi,\varphi}(\vec{x}, \vec{y}) = \min\{\pi^{\nu}(\vec{x} - A\vec{y}), \pi^{\varphi}(\vec{y})\}$, действуя так же, как описано во введении при выводе формулы (6).

Далее, так как $\vec{u} = U\vec{\varphi}$ с возможностью единица, и нарушение этого равенства невозможно, то $\pi^{\xi,\varphi,u}(\vec{x}, \vec{y}, \vec{z}) = \pi^{\xi,\varphi}(\vec{x}, \vec{y})$, если $\vec{u} = U\vec{\varphi}$, и $\pi^{\xi,\varphi,u}(\vec{x}, \vec{y}, \vec{z}) = 0$, если $\vec{u} \neq U\vec{\varphi}$, то совместное распределение нечетких векторов $\vec{u}$ и $\vec{\xi}$ есть

$$\pi^{\xi,u}(\vec{x}, \vec{z}) = \sup_{\vec{y}} \begin{cases} \min\{\pi_0(\|\vec{x} - A\vec{y}\|_{\varepsilon}), \pi^{\varphi}(\vec{y})\}, & \text{если } \vec{z} = U\vec{y}, \\ 0, & \text{в противном случае.} \end{cases}$$

Заметим, что максимальное по значение функции $\pi^{\xi,u}(\vec{x},\cdot)$, т.е. решение задачи $\vec{z}_*(\vec{x}) = \arg\max\limits_{\vec{z}\in\mathcal{R}^N}\{\pi^{\xi,u}(\vec{x},\vec{z})\}$, достигается при $\vec{z}_*(\vec{x}) = U\vec{y}_*(\vec{x})$, где $\vec{y}_*(\vec{x})$ — решение задачи

$$\vec{y}_*(\vec{x}) = \arg\max_{\vec{y}\in\mathcal{R}^n}\{\min\{\pi_0(\|\vec{x}-A\vec{y}\|_\varepsilon),\pi^\varphi(\vec{y})\}\}. \tag{19}$$

Обозначим $\vec{d}(\cdot) : \mathcal{R}_n \to \mathcal{R}^N$ функцию, для каждого значения $\vec{\xi} = \vec{x}$ определяющую оценку $\hat{\vec{u}} = d(\vec{x})$ вектора $\vec{u}$, и $l(\cdot,\cdot)$: $\mathcal{R}^n \times \mathcal{R}^n \to [0,1]$ — распределение возможностей потерь: значение $l(\vec{d},\vec{u})$ есть возможность потерь в ситуации, когда вместо значения $\vec{u} = \vec{z}$ используется его значение $\vec{d}$. Тогда, как описано во введении, возможность потерь при оценке $\vec{d}(\cdot)$ и при заданном значении измерения $\vec{\xi} = \vec{x}$ равна $PL(\vec{d}(\vec{x})) = \sup\limits_{\vec{y}}\min\{l(\vec{d}(\vec{x}),\vec{y}),\pi^{\xi,u}(\vec{x},\vec{y})\}$.

Оценка $\vec{d}_*(\cdot)$, минимизирующая максимальную возможность потерь, для каждого $\vec{x}\in\mathcal{R}^n$ определяется решением задачи на минимум $PL(\vec{d}_*(\vec{x})) = \min\limits_{\vec{d}(\cdot)} PL(\vec{d}(\vec{x}))$.

Как показано в [15], если функция возможности потерь задана так, что нулевая возможность потерь соответствуют верному решению, а любое неверное решение влечет потери с возможностью единица, то оценку максимальной возможности потерь определяет значение $\vec{d}_*(\vec{x})$, которое доставляет максимум возможности $\pi^{\xi,u}(\vec{x},\vec{y})$ для каждого $\vec{\xi} = \vec{x}$. Такие оценки называются оценками максимальной апостериорной возможности. Таким образом, задача построения оценки, минимизирующей возможность потерь, свелась к задаче на максимум (19), и для ее решения следует найти максимум $\vec{y}_*(\vec{x})$ функции

$$\min\{\pi_0(\|\vec{x}-A\vec{y}\|_\varepsilon),\pi^\varphi(\vec{y})\}. \tag{20}$$

и вычислить оценку $\vec{d}_*(\vec{x}) = U\vec{y}_*(\vec{x})$.

В силу монотонности функции $\pi_0(\cdot)$ задача на максимин (20) эквивалентна задаче на минимакс

$$\min_{\vec{y}:(\vec{w}_m,\vec{y})_n\in[l_m,r_m],m=1,\ldots,n}\left\{\max_{i=1,\ldots,n}\frac{\left|x_i-\sum\limits_{k=1}^n A_{ik}y_k\right|}{\varepsilon_i}\right\}. \tag{21}$$

Задача (21) сводится к задаче линейного программирования [13, 21] и по сути является модифицированной задачей минимизации невязки при решении системы линейных алгебраических уравнений. Заметим, что если значение минимума функционала в (21) больше единицы, то возможность такой оценки в силу (18) равна нулю, что свидетельствует о неадекватности используемой математической модели.

Описание и свойства оценки максимальной возможности суммируем в следующем утверждении.

Теорема. *Пусть в схеме измерений* (9) $a_i(\cdot)$, $i = 1,\ldots,n$ *— непрерывные функции, квадрат которых интегрируем на* $[0,\infty)$, ν_i, $i = 1,\ldots,n$, *— независимые нечеткие элементы с распределением возможностей* (18). *Тогда оценкой максимальной возможности значений ортогональной проекции* $Pg(\cdot)$ *функции* $g(\cdot)$ *на линейную оболочку* S_A *функций*

$a_i(\cdot)$, $i = 1, \ldots, n$, в точках $\{\lambda_1, \ldots, \lambda_N\} \in [0, \infty)$, удовлетворяющих условиям (12), (14) *является* $P\hat{g}(x_i) = \sum\limits_{k=1}^{n} y_{*j} a_k(x_j)$, $i = 1, \ldots, N$, *где* y_{*i}, $i = 1, \ldots, N$, — *решение задачи* (21). *Если значение минимума функционала в* (21) *больше единицы, то математическая модель измерения* (9) *не согласуется с его результатом.*

Оценки максимальной возможности параметров спектрометрического эксперимента

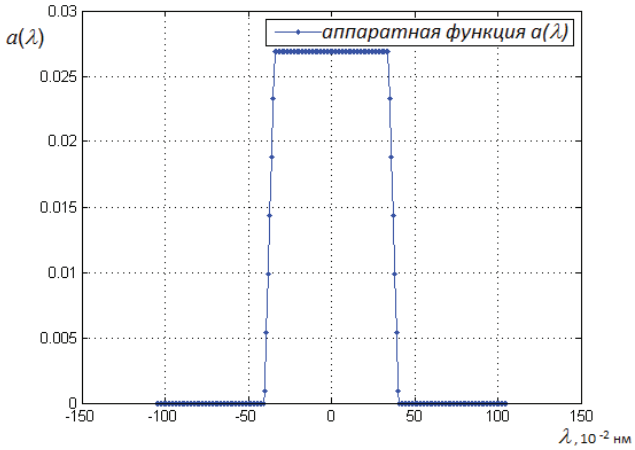


Рис. 2. График аппаратной функции двухщелевого спектрометра $a(\cdot)$, усл.ед.

Эффективность методов, предложенных в предыдущих разделах статьи, применялись для интерпретации результатов измерений интенсивности излучения газоразрядной лампы на длинах волн в окрестности эмиссионного спектра D-линии натрия (дублета с положением максимумов $\lambda_1 = 588.9950$ нм и $\lambda_2 = 589.5924$ нм). Измерения проводились на двулучевом спектрометре, ширины щелей спектрометра были выбраны так, чтобы аппаратная функция имела трапециевидную форму [1]. График аппаратной функции на интервале от -1 нм до 1 нм при фиксированных d_1 и d_2 приведен на рисунке 2, один отсчет по горизонтальной оси соответствует 0.01 нм. По вертикальной оси — условные единицы.

Спектр измерялся согласно схеме измерения $\xi(\lambda_i) = \int\limits_{-\infty}^{\infty} a(\lambda - \lambda_i) f(\lambda) d\lambda + \nu_i$, $i = 1, \ldots, 211$, в которой длины волн $\lambda_1, \ldots, \lambda_{211}$ изменялись от 588.20 нм до 590.30 нм с шагом 0.01 нм. Результат измерения спектра — вектор с координатами $(\xi(\lambda_1), \ldots, \xi(\lambda_{211}))$ — изображен на рисунке 3. Здесь по горизонтальной оси отложены значения двухсот одиннадцати длин волн, в которых измерялся спектр, по вертикальной оси отложена интенсивность измеренного спектра, равная числу зарегистрированных фотонов. Хотя экспериментальные данные были получены в результате реальных измерений, целью эксперимента была демонстрация возможностей математических методов интерпретации измерений, поэтому ширина аппаратной функции была специально выбрана существенно большей, чем расстояние между линиями дублета, в измеренном спектре они не разрешены.

Значение длин волн, в которых оценивалось значение входного спектра, были выбраны на том же отрезке от 588.20 до 590.30 нм с шагом 0.01 нм, таким образом, оцениваемый вектор имеет размерность 211. Считалось, что значение искомой проекции для всех 211 длинах волн неотрицательны, то есть в (16) $u_{i,\min} = 0$, $u_{i,\max} = +\infty$, $i = 1, \ldots, 211$.

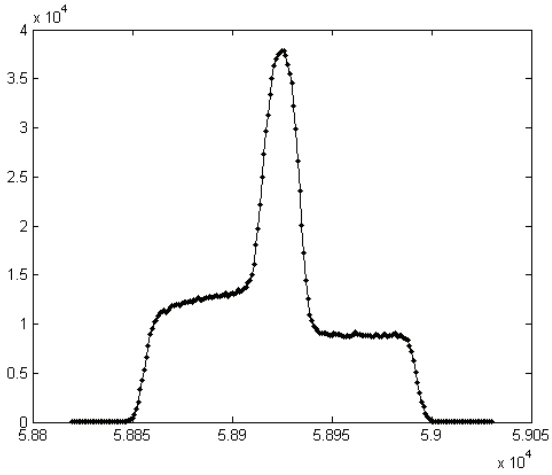


Рис. 3. Результат измерения спектра излучения Na: число зарегистрированных фотонов на каждой длине волны $\lambda_1, \ldots, \lambda_{211}$

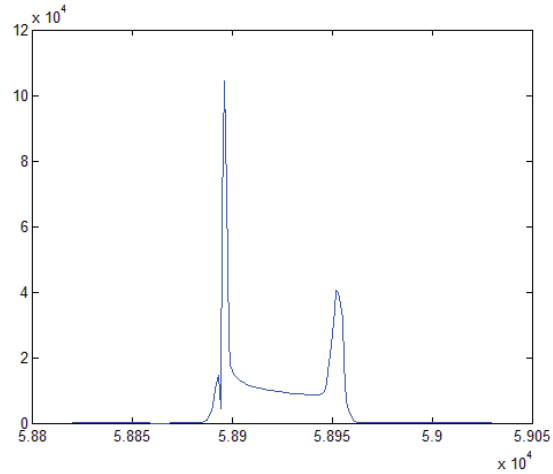


Рис. 4. Оценка входного спектра методом максимальной возможности, усл.ед.

Величина максимальной погрешности измерений ε_i для каждой длины волны λ_i считалась пропорциональной квадратному корню из ξ_i, $i = 1, \ldots, 211$. Априорные ограничения на значения оцениваемого спектра состояли только в требовании неотрицательности.

Оценка значений проекции спектра на S_A в выбранных значениях длин волн, минимизирующая максимальную возможность потерь, приведена на рисунке 4. Видно, что вблизи максимумов дублета при $\lambda_1 = 588.9950$ нм и $\lambda_2 = 589.5924$ нм в оценке спектра имеются существенные максимумы.

Минимаксная оценка значений проекции спектра при тех же значениях длин волн приведена на рисунке 5. Верхняя кривая есть график $u_{j,\max}$ правой границы интервалов возможных значений оцениваемой координаты u_j, $i = 1, \ldots, 211$. Нижний график (он практически совпадает с осью абсцисс) есть график левой границы $u_{j,\min}$ интервалов возможных значений оцениваемой координаты u_j, $j = 1, \ldots, 211$. График свидетельствует, что практически для каждой координаты, рассматриваемой независимо от остальных координат, возможно ее значение, равное нулю. Минимаксная оценка координат u_j дается графиком, расположенным посередине между значениями $u_{j,\min}$ и $u_{j,\max}$. Из рисунка

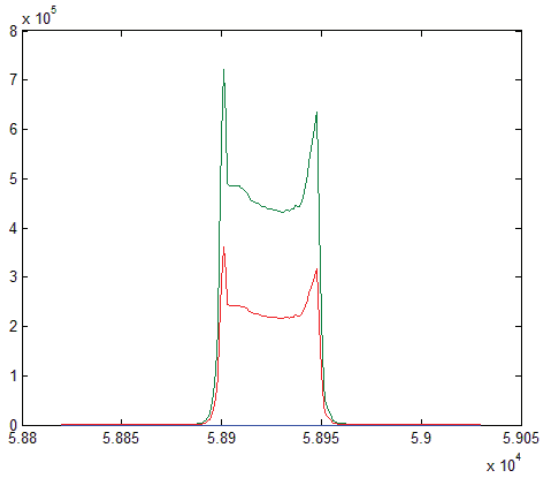


Рис. 5. Минимаксная оценка входного спектра, усл.ед.

видно, что в данных спектрометрического эксперимента содержится недостаточно информации для того, чтобы уверенно обнаружить дублет излучения Na.

Учет же дополнительной информации о том, что большие ошибки измерений менее возможны, чем малые, приводит к существенно лучшему результату, см. рисунок 4.

Заключение

В работе изучена задача оценивания входного сигнала измерительного прибора на основе математической модели, описывающей эксперимент, в котором измеряется конечное число линейных функционалов от входного сигнала прибора. Измерения сопровождаются погрешностью. Измеряемые в эксперименте линейные функционалы определяются как значение свертки входного сигнала с аппаратной функцией измерительного прибора; аппаратная функция считается непрерывной на своей области определения. Показано, что с конечной погрешностью можно оценить лишь конечномерную составляющую входного сигнала. Построены оценки, минимизирующие максимальную погрешность оценивания искомых значений, при этом используется априорная информация об оцениваемых величинах, заданная системой линейных неравенств, и предполагается, что измерительная погрешность может быть любой в пределах заданного интервала. Показано, что предположение о том, что большие погрешности измерений менее возможны, чем малые, могут существенно улучшить результат оценивания. Для формализации априорных предположений о том, что большие погрешности менее возможны, чем малые, построена возможностная модель измерения, поставлена и решена задача оценивания путем минимизации максимальной возможности потерь.

Эффективность методов иллюстрируются оценкой значений оптического спектра по данным, полученным с двухлучевого спектрометра с широкими ширинами щелей. Показано, что предположение о большей возможности малых значениях погрешности измерений позволяет получить более адекватное представление об измеряемом спектре, чем при оценивании, минимизирующем максимальную погрешность оценки.

Литература

[1] *Лебедева В. В.* Экспериментальная оптика. 4-е изд. М.: Физический факультет, МГУ им. М.В.Ломоносова, 2005. 282 с.

[2] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. М.: Наука, 1979.

[3] *Федотов А. М.* Некорректные задачи со случайными ошибками в данных. Новосибирск: Наука, 1990.

[4] *Турчин В. А., Козлов В. П., Малкевич М. С.* Использование методов математической статистики для решения некорректных задач // *УФН*, 1970. Т. 102, № 3. С. 345–385.

[5] *Иванов В. К., Васин В. В., Танана В. П.* Теория линейных некорректных задач и ее приложения. М.: Наука, 1978.

[6] *Лаврентьев М. М.* О некоторых некорректных задачах математической физики. Новосибирск: Изд-во СО АН СССР, 1962.

[7] *Тихонов А. Н., Гончарский А. В., Степанов В. В., Ягола А. Г.* Численные методы решения некорректных задач. М.: Наука, 1990. 232 с.

[8] *Ягола А. Г., Ван Янфей, Степанова И. Э., Титаренко В. Н.* Обратные задачи и методы их решения. Приложения к геофизике. М.: Бином. Лаборатория знаний, 2014.

[9] *Самарский А. А., Вабищевич П. Н.* Численные методы решения обратных задач математической физики. М.: Издательство ЛКИ, 2009. 480 с.

[10] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.

[11] *Eichstadt S., Schmahling F., Wubbeler G., Kruger U., Elster C.* A new approach to bandpass correction in spectrometer measurements using the Richardson-Lucy method // *16-th International Congress of Metrology*, 2013. 14005.

[12] *Silva Neto A. J., Cella N.* A regularized solution with weighted Bregman distances for the inverse problem of photoacoustic spectroscopy // *Comput. Appl. Math*, 2006. Vol.25. P. 139–165.

[13] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. М.: ФИЗМАТЛИТ, 2012. 428 с.

[14] *Лидов М. Л.* Минимаксные методы оценивания. М.: Препринты ИПМ им. М.В.Келдыша. № 71, 2010. 87 с.

[15] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применение. М.: ФИЗМАТЛИТ, 2007. 464 с.

[16] *Zadeh L.* Fuzzy Sets // *Information and Control*, 1965. Vol. 8. P. 235–350.

[17] *Zadeh L.* Fuzzy sets as a basis for a theory of possibility // *Fuzzy Sets and Systems*, 1999. No. 100. P. 9–24.

[18] *Dubois D., Prade H.* Possibility Theory, Probability Theory and Multiple-valued Logics: A Clarification // *Annals of Mathematics and Artificial Intelligence*, 2001. Vol. 32. P. 35–66.

[19] *Dubois D., Prade H.* Formal representation of uncertainty // *Decision-Making Process* / Ed. by D. Bouyssou, D. Dubous, V. Pirlot, H. Prade. ISTE, London, 2009. P. 85–156.

[20] *Yager R.* Conditional Approach to Possibility-Probability Fusion // *IEEE Transactions on Fuzzy Systems*, 2012. Vol. 20, No. 3. P. 526–535.

[21] *Кириллов К.В., Чуличков А. И.* Редукция измерений в нечеткой модели эксперимента как решение задачи линейного программирования // *Вестн. Моск. ун-та. Физ. Астрон*, 1999. № 2. С. 62–64.

References

[1] *Lebedeva V. V.* 2005. Experimental optics. 4th ed. M.: Faculty of Physics, Moscow State University. 282 p. (in Russ.)

[2] *Tikhonov A. N., Arsenin V. Y.* 1979. Methods for solving ill-posed problems. M.: Nauka. (in Russ.)

[3] *Fedotov A. M.* 1990. Ill-posed problems with random errors in the data. Novosibirsk: Nauka. (in Russ.)

[4] *Turchin V. A., Kozlov V. P., Malkevich M. C.* 1970. The use of mathematical statistics methods for solving ill-posed problems. *Physics-Uspekhi (Advances in Physical Sciences)* 102(3):345–385. (in Russ.)

[5] *Ivanov V. K., Vasin V. V., Tanana V. P.* 1978. The theory of linear ill-posed problems and its applications. M.: Nauka. (in Russ.)

[6] *Lavrent'ev M. M.* 1962. On some ill-posed problems of mathematical physics. Novosibirsk: SD AS USSR Publ. (in Russ.)

[7] *Tikhonov A. N., Goncharsky A. V., Ctepanov V. V., Yagola A. G.* 1990. Numerical methods for solving ill-posed problems. M.: Nauka. 232 p. (in Russ.)

[8] *Yagola A. G., Van Yunfey, Stepanova I. E., Titarenko V. N.* 2014. Inverse problems and methods of their solution. Applications to Geophysics. M.: Binom. Laboratoria znanij. (in Russ.)

[9] *Samarsky A. A., Vabishevich P. N.* 2009. Numerical methods for solving inverse problems of mathematical physics. M.: LKI Publisher. 480 p. (in Russ.)

[10] *Vapnik V. N.* 1979. Restore dependency on empirical data. M.: Наука. (in Russ.)

[11] *Eichstadt S., Schmahling F., Wubbeler G., Kruger U., Elster C.* 2013. A new approach to bandpass correction in spectrometer measurements using the Richardson-Lucy method. *16-th International Congress of Metrology.* 14005.

[12] *Silva Neto A. J., Cella N.* 2006. A regularized solution with weighted Bregman distances for the inverse problem of photoacoustic spectroscopy. *Comput. Appl. Math* 25:139–165.

[13] *Pyt'ev Yu. P.* 2012. Mathematical modeling methods of measuring computer-aided systems. M.: FIZMATLIT. 428 p. (in Russ.)

[14] *Lidov M. L.* 2010. Minimax estimation methods. M.: Preprint IPM. No. 71. 87 p. (in Russ.)

[15] *Pyt'ev Yu. P.* 2007. The possibility as an alternative to probability. Mathematical and empirical basis, application. M.: FIZMATLIT. 464 p. (in Russ.)

[16] *Zadeh L.* 1965. Fuzzy Sets. *Information and Control* 8:235–350.

[17] *Zadeh L.* 1999. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 100:9–24.

[18] *Dubois D., Prade H.* 2001. Possibility Theory, Probability Theory and Multiple-valued Logics: A Clarification. *Annals of Mathematics and Artificial Intelligence* 32:35–66.

[19] *Dubois D., Prade H.* 2009. Formal representation of uncertainty. In: Bouyssou D., Dubois D., Pirlot M., Prade H. (eds) Decision-making process, ISTE, London. 85–156.

[20] *Yager R.* 2012. Conditional Approach to Possibility-Probability Fusion. *IEEE Transactions on Fuzzy Systems* 20(3):526–535.

[21] *Kirillov K. V., Chulichkov A. I.* 1999. The reduction of measurements in fuzzy model of the experiment as the solution of a linear programming problem. *Moscow University Physics Bulletin* 2:62–64. (in Russ.)

Классификация видов физической активности человека по показаниям акселерометра и гироскопа

О. А. Харацидиди
oleg.kharatsidi@gmail.com
Московский государственный университет им. М. В. Ломоносова

Рассматривается задача распознавания видов физической активности человека по показаниям акселерометра и гироскопа портативного устройства на примере открытого набора данных USC-HAD с 12 классами. Предлагается метод, использующий иерархию классов и настраивающий отдельные классификаторы в ее узлах. Ключевую роль играют классификаторы, использующие частотные признаки и представляющие собой смесь из трех принципиально различных моделей: логистической регрессии, метода ближайшего соседа и случайного леса. Итоговое качество классификации соответствует среднему значению F-меры 0,92.

Ключевые слова: распознавание физической активности; сенсоры; обработка сигналов

Human activity recognition based on accelerometer and gyro data

O. A. Kharatsidi
M. V. Lomonosov Moscow State University

In the last few years the performance of smartphones has been growing rapidly. They have become capable of carrying out relatively complex computations in real time. In the paper, a problem of human physical activity recognition is considered based on the data from sensors on wearable devices. The classical approach is to split the accelerometer signals into fixed-width windows, classify them independently, and then combine the classification responses into a single one for the whole sample. Classification models may vary from Naïve Bayes classifiers to Neural Nets. The two common types of the features are the statistical metrics (moments, correlations, etc.) and Fourier coefficients.

The method introduced in this paper utilizes the same approach but unlike the most common case, it uses a mixture of three standard classification models: Logistic Regression, Nearest Neighbour, and Random Forest built up on the Fourier coefficients absolute values. Feature selection based on the trained Logistic Regression coefficients is applied to fit the rest two models independently. The method was tested on the USC-HAD open dataset containing measurements of 12 classes from 14 people. Apart from the widely used accelerometer data, it also provides gyro signals which are used in just the same way. The method also exploits a hierarchy of the classes and trains multiple individual classifiers in its nodes.

Since the data consists of the measurements for multiple people, in their experiments, the authors run cross-validation with a single fold per each person. In each iteration, an internal cross-validation was also run to fit hyperparameter. As a result, the algorithm achieves the performance of 0.92 in terms of the mean F-measure. The experiments also show that the mixture of the three models is more stable than each of its components and achieves higher performance. Finally, the method proves to be significantly better than standard L_2-regularized Logistic Regression built up on the same feature set.

Keywords: activity recognition; sensors; signal processing; context recognition

Введение

За последние несколько лет возможности смартфонов значительно выросли, а их использование стало повсеместным. Появились целые рынки приложений для мобильных операционных систем. На этом фоне возросла потребность в решении целого класса задач (context recognition [1]) по определению местоположения, активности или состояния пользователя, а также анализу его физической деятельности с помощью датчиков устройства: акселерометра, гироскопа, Bluetooth-адаптера, микрофона, датчиков освещения, давления и других. Росту интереса к таким задачам также способствует появление носимой электроники (wearable electronics) — нового класса устройств, таких как, например, «умные часы».

Один из простых примеров конкретных задач, решаемых с помощью показаний датчиков, — подсчет количества шагов пользователя.

В данной работа рассматривается другая известная задача — распознавание видов физической активности пользователя (activity recognition [2, 3, 4, 5, 6, 7]). В качестве исходных данных берутся показания акселерометра и гироскопа (датчика угловой скорости).

Описание данных

В данной работе рассматривается задача классификации для набора данных USC-HAD [8], находящегося в открытом доступе.

Показания акселерометра и гироскопа (датчика угловой скорости) снимались с прибора, крепящегося в районе пояса в определенном положении, с частотой 100 Гц. Всего участвовало 14 чел., каждый из которых сделал по 5 подходов на каждый из 12 видов физической активности. Каждый подход длился примерно от 10 до 30 с.

Показания снимались в моменты, когда человек:

(1) идет вперед;
(2) идет по кругу, поворачивая влево;
(3) идет по кругу, поворачивая вправо;
(4) поднимается по лестнице;
(5) спускается по лестнице;
(6) бежит вперед;
(7) прыгает на месте;
(8) сидит с небольшими движениями;
(9) стоит на месте;
(10) лежит;
(11) поднимается на лифте;
(12) спускается на лифте.

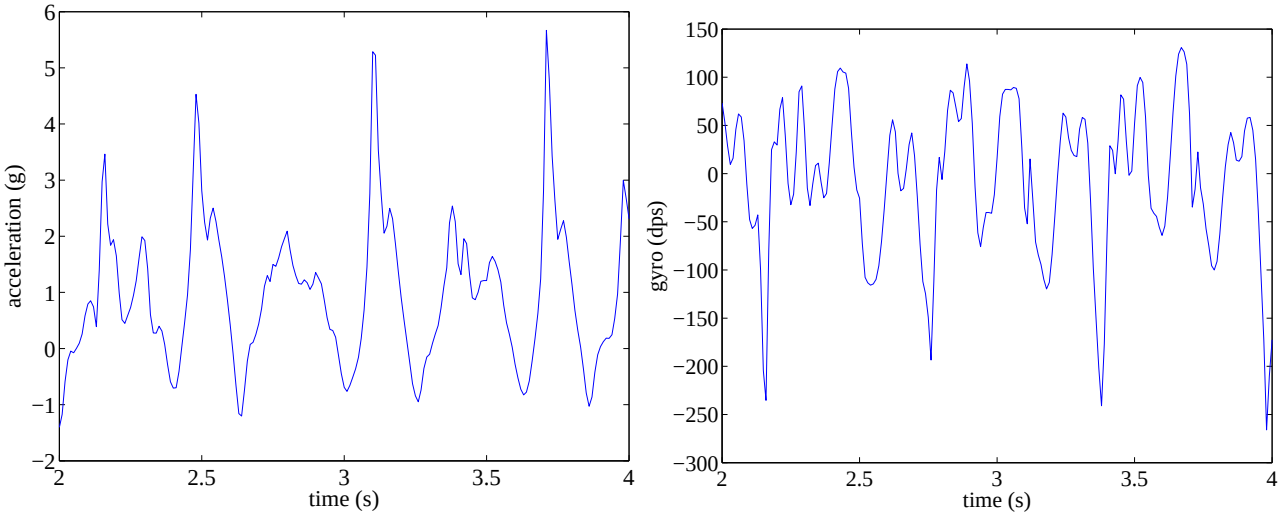
Каждому из приведенных видов активности соответствует класс. В качестве объекта рассматриваются показания 3-х осей акселерометра и 3-х осей гироскопа (рис. 1).

Таким образом, в выборке 12 классов по $5 \cdot 14 = 70$ объектов в каждом.

Стоит отметить, что с этим набором данных работать несколько проще, чем с показаниями датчиков современных телефонов: качество и характеристики датчиков телефонов сильно разнятся, и, самое главное, телефоны не носят в каком-то определенном положении.

С другой стороны, на этом наборе данных легко переобучиться, поэтому в данной работе по возможности избегается тонкая настройка гиперпараметров.

Более подробное описание данных и способа их получения можно найти в [8].



(а) (б)

Рис. 1. Фрагмент показаний одной из осей акселерометра (а) и гироскопа (б) для класса «бег»

Метод решения

Прежде чем перейти непосредственно к описанию моделей, необходимо упомянуть несколько достаточно стандартных идей, нашедших применение также и в предложенном в этой работе методе:

— Для каждого объекта рассматриваются окна шириной в 512 отсчетов, соседние перекрываются на 50%. Каждое такое окно, содержащее по $512 \cdot 6$ значений, будем рассматривать как отдельный объект одного из 12 исходных классов.
Строго говоря, это уже другая задача классификации, но используя ее результат, можно получить решение для *исходной* (далее везде будет использоваться именно такое название), если построить решающее правило по принципу голосования: с учетом результатов классификации окон строится множество голосов для данного объекта исходной тестовой выборки, затем этот объект относится к тому классу, за который отдано наибольшее количество голосов, т. е.

$$a(x) = \arg\max_{y \in Y} |\{v \in \mathcal{V}(x) | v = y\}|. \tag{1}$$

Здесь и далее x — объект *исходной* задачи классификации; Y — множество меток классов; $\mathcal{V}(x)$ — множество (формально — мультимножество) всех голосов, соответствующих объекту x. Классификаторы для *исходной* задачи будут обозначаться $a(\cdot)$, а классификаторы для окон — $\tilde{a}(\cdot)$. Множество $\mathcal{V}(x)$ будет в разных случаях вводиться несколько по-разному с использованием множества $\mathcal{W}(x)$ окон объекта x;

— по аналогии с [6], все классы делятся на две группы: *активные* и *пассивные*. Классы 1–7 — активные, классы 8–12 — пассивные. Классификатор строится по следующему принципу: сначала тестовый объект относится одним классификатором к одной из двух групп, а затем подается на вход другому алгоритму для классификации внутри данной группы.
Эффективность такого подхода заключается в том, что, как будет показано далее, пассивные и активные действия хорошо разделяются с использованием очень простых

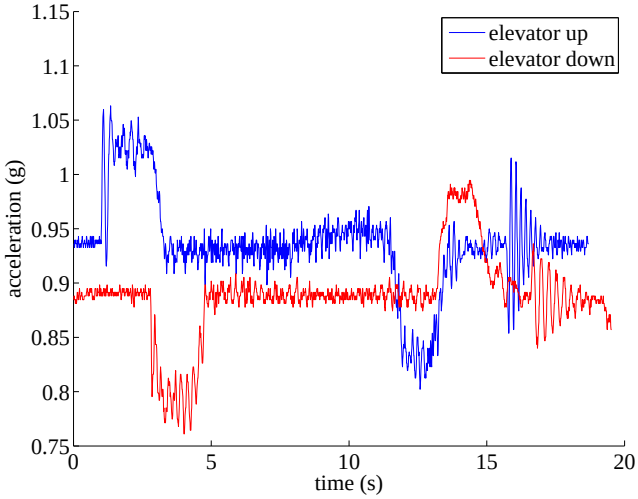


Рис. 2. Пример: передвижение на лифте вверх и вниз, ускорение по вертикали

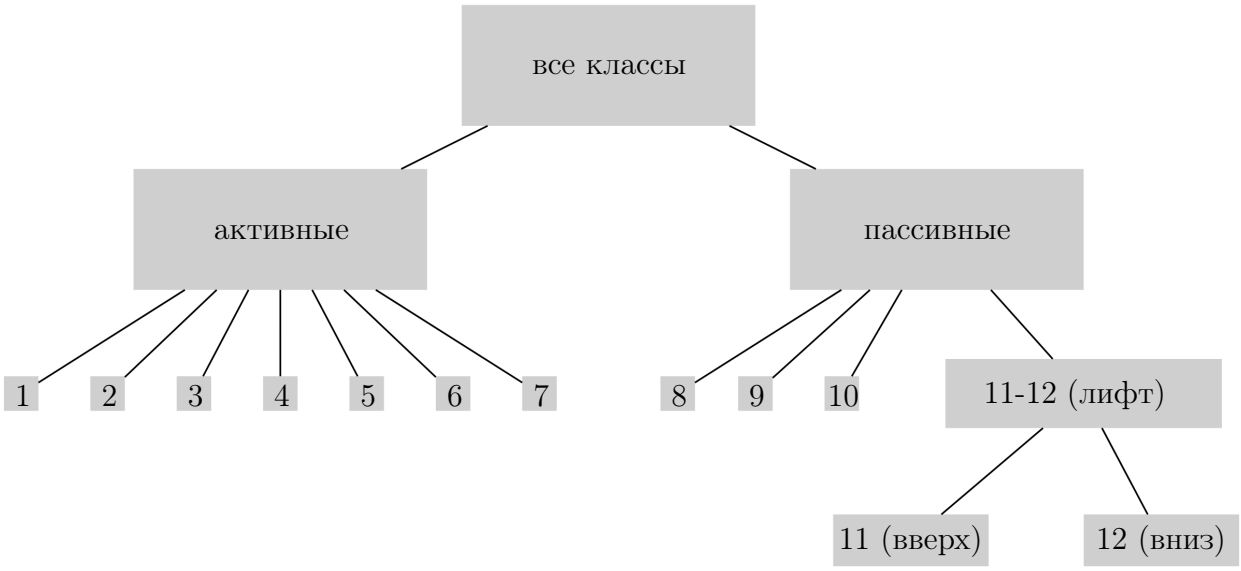


Рис. 3. Иерархия классов

признаков, в то время как для дальнейшей классификации удобно использовать признаки совсем иной природы;

— необходимо отметить, что подход с выделением окон и голосованием не применим для разделения классов 11 и 12 (передвижение на лифте вверх и вниз). Сигналы этих классов имеют характерные «всплески», порядок которых необходимо учесть (рис. 2). Поэтому изначально при обучении и выводе эти классы рассматриваются как один, и только на последней стадии разделяются отдельным классификатором, не использующим окна.

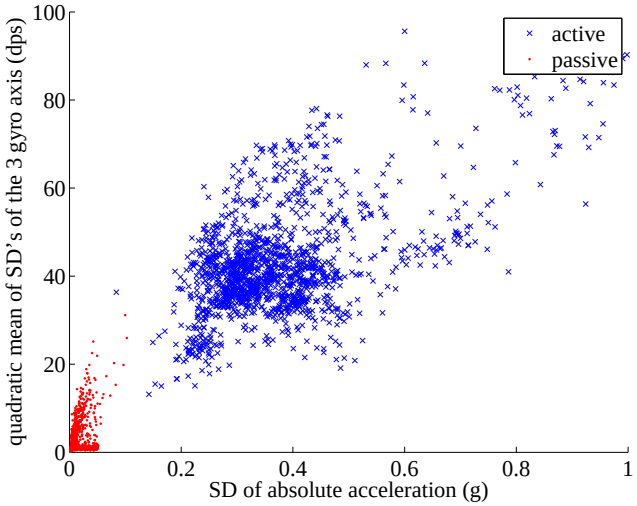


Рис. 4. Объекты, соответствующие активным и пассивным группам классов

В пояснение к последним двум идеям на рис. 3 изображена иерархия классов, в соответствии с которой производится классификация. Каждой нелистовой вершине дерева соответствует свой классификатор.

Голосование происходит на каждом отдельном этапе, кроме этапа разделения классов 11 и 12, на котором используются не окна, а объекты исходной выборки.

Наибольший интерес представляют алгоритмы классификации, соответствующие узлам всех активных и всех пассивных классов. Другие два классификатора очень просты, и рассматриваются как вспомогательные.

Разделение активных и пассивных действий

В вершине предложенной иерархии происходит разделение классов на активные и пассивные. Классификатор, используемый для такого разделения, устроен очень просто.

Для каждого окна можно вычислить стандартное отклонение (standard deviation, SD) абсолютного значения ускорения и среднее квадратичное стандартных отклонений трех каналов угловых скоростей. Эти два признака в некоторой мере отражают «степень активности», фиксируемую акселерометром и гироскопом, соответственно.

На рис. 4 видно, что выборка хорошо разделима по первому признаку. Второй же признак выглядит менее информативным, поэтому использоваться не будет.

Далее можно воспользоваться пороговым классификатором по первому признаку

$$\tilde{a}(w) = [f(w) - b > 0] = \begin{cases} 1, & \text{если } f(w) - b > 0; \\ 0, & \text{иначе} \end{cases}$$

с функцией потерь

$$\ell(M) = (-M)_+ = \begin{cases} -M, & \text{если } M < 0; \\ 0, & \text{иначе,} \end{cases}$$

где $M(x)$ — отступ объекта x; w — окно; $f(w)$ — признак окна w (среднее квадратичное стандартных отклонений).

Такой классификатор имеет единственный настраиваемый параметр b — значение порога. Средняя функция потерь на всей выборке может иметь плато, поэтому в качестве

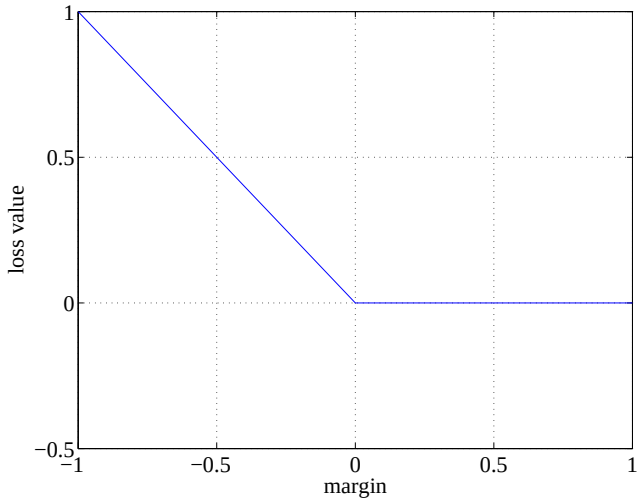


Рис. 5. Функция потерь $\ell(M) = (-M)_+$

порога берется среднее арифметическое минимального и максимального оптимальных значений.

Голосование строится в соответствии с (1), множество голосов $\mathcal{V}(x)$ определяется следующим образом:

$$\mathcal{V}(x) = \{\tilde{a}(w) | w \in \mathcal{W}(x)\},$$

где $\mathcal{W}(x)$ — множество окон объекта x; $\tilde{a}(\cdot)$ — классификатор окон.

Несмотря на примитивность, этот метод показывает качество 100% в рамках исходной задачи классификации.

Классификация передвижений на лифте

Как видно на рис. 2, для движения лифта вверх характерен «скачок» ускорения по вертикали вначале и «провал» в конце, а для движения вниз — наоборот. Поэтому самый простой способ разделить эти два класса — посчитать среднее ускорение по вертикали в первой половине всего временного интервала и вычесть из него среднее ускорение по вертикали во второй половине.

Таким образом, в данном случае классификатор описывается следующим образом:

$$\tilde{a}(w) = \begin{cases} 0, & \text{если } m_1(x) < m_2(x); \\ 1, & \text{иначе,} \end{cases}$$

где $m_1(x)$ и $m_2(x)$ — среднее ускорение по вертикали в первой и во второй половинах временного интервала, соответственно.

Как видно по графику на рис. 6, такой способ дает почти идеальное качество классификации. Кроме того, получившийся классификатор не требует обучения и имеет вполне естественный логический смысл.

Основная модель

Для классификации внутри групп активных и пассивных классов используется общая модель, но обучение происходит отдельно. При этом в группе пассивных действий классы, соответствующие передвижению на лифте вверх и вниз, на данном этапе объединены.

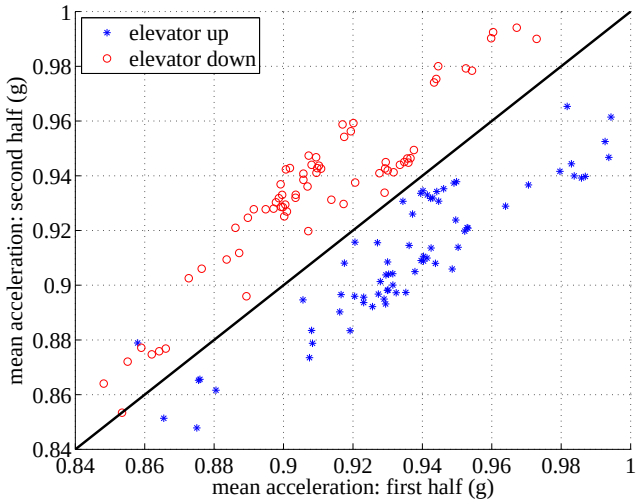


Рис. 6. Движение лифта вверх и вниз — разделимость классов

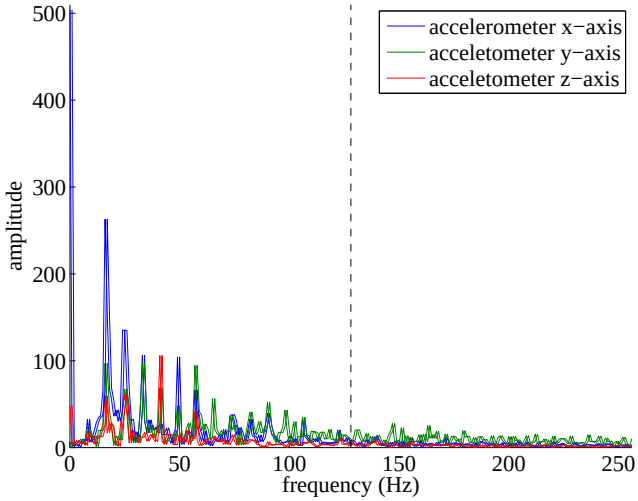


Рис. 7. Амплитудный спектр для показаний акселерометра. Для классификации используются только значения из левой половины графика

В качестве признакового пространства используются низкочастотные половины (128 значений) амплитудных спектров всех шести сигналов, полученные с помощью преобразования Фурье (рис. 7).

Каждый признак центрируется и нормируется на квадратный корень из нормы.

Модель состоит из смеси логистической регрессии, метода ближайшего соседа (Nearest Neighbor) и случайного леса (Random Forest), обозначаемых $\tilde{a}_1(\cdot)$, $\tilde{a}_2(\cdot)$ и $\tilde{a}_3(\cdot)$ соответственно. При этом с помощью логистической регрессии также производится отбор признаков, используемый для каждой из двух других подмоделей.

Далее приведено более подробное описание всех подмоделей.

1. **Логистическая регрессия с L_1-регуляризацией.** Естественно предположить, что спектр содержит много избыточной информации, поэтому имеет смысл рассмотреть линейную логистическую регрессию с L_1-регуляризацией. В данной работе используется реализация из библиотеки LIBLINEAR [9] с интерфейсом для системы MATLAB.

Это обычная двухклассовая логистическая с добавлением единичного константного признака (bias), обучаемая по принципу «один против всех».

Единственный настраиваемый гиперпараметр — значение C, обратное коэффициенту регуляризации, — настраивается на скользящем контроле.

С использованием коэффициентов обученной логистической регрессии вводится понятие значимости признаков.

Пусть $p(\boldsymbol{x}) = (1 + \exp(-W^{\intercal}\boldsymbol{x}))^{-1}$ — вектор оценок вероятности принадлежности объекта $\boldsymbol{x}$ к классам (ненормированный); W — настроенные веса, каждый столбец соответствует своему классу. Тогда мера значимости i-го признака будет определяться как $s_i = max_j|W_{ij}|$.

2. **Nearest Neighbor.** В качестве следующей подмодели берется обычный метод одного ближайшего соседа с косинусной мерой. В силу того, что признаки уже нормализованы, косинусная мера в данном случае эквивалентна корреляции. Из признаков используется только половина (т. е. 64) наиболее значимых. В качестве ответа возвращается класс самого ближайшего по косинусной мере объекта из обучающей выборки.
3. **Random Forest.** Используется реализация бэггинга над решающими деревьями из пакета Statistics Toolbox системы MATLAB. Количество деревьев — 1000, для остальных параметров используются значения по умолчанию.

   Здесь, по аналогии с NN, используется только половина наиболее значимых признаков.

Все три подмодели обучаются независимо, если не считать отбор признаков. В рамках исходной задачи, на стадии вывода для каждого объекта тестовой выборки «голоса» моделей объединяются, после чего принимается решение о классификации данного объекта по формуле (1). Формально $V(x)$ в данном случае определяется так:

$$\mathcal{V}(x) = \mathcal{V}_1(x) \cup \mathcal{V}_2(x) \cup \mathcal{V}_3(x),$$

где $\mathcal{V}_i(x) = \{\tilde{a}_i(w)|w \in \mathcal{W}(x)\}$.

Такая композиция — не что иное, как усреднение трех различных подмоделей.

Эксперименты

Способ оценки качества

Прогноз модели делается на скользящем контроле, затем итоговое качество вычисляется как усредненная F-мера по всем классам. *Каждый фолд скользящего контроля соответствует отдельному человеку.* Таким образом, всего имеется 14 фолдов: в i-й входят данные, соответствующие i-му человеку.

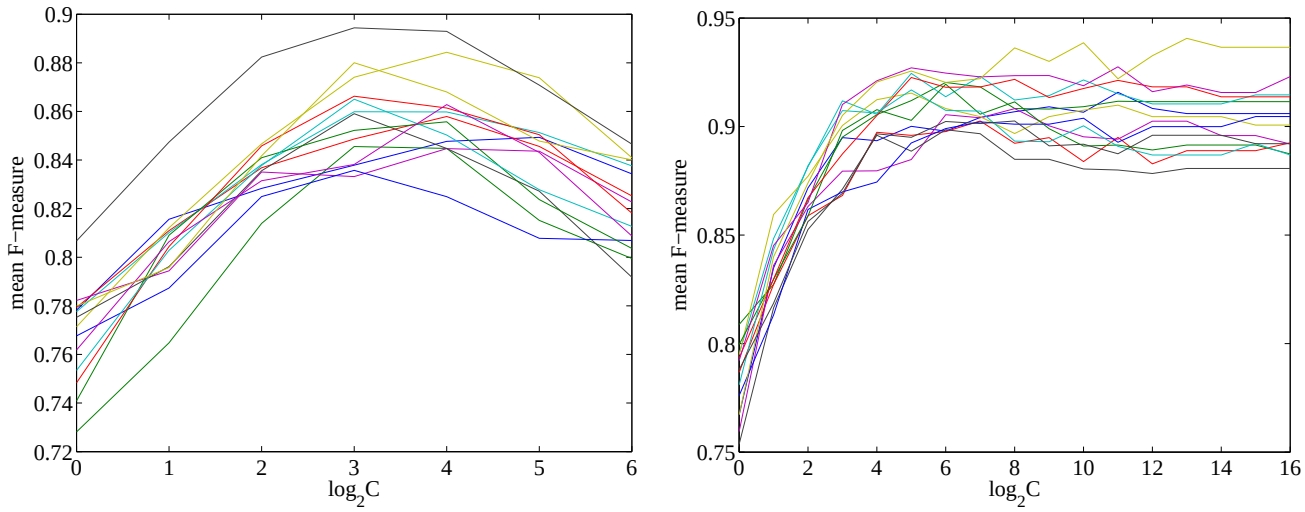
Параметры нормировки признаков (средние значения и нормы) вычисляются на каждой итерации скользящего контроля по текущей обучающей подвыборке.

Везде оценивается качество в рамках исходной задачи классификации.

Результаты

Далее изложены результаты экспериментов для отдельных классификаторов, для подмоделей и для всей композиции. Некоторые из этих результатов уже упоминались выше. Также приводится сравнение с простой логистической регрессией.

— классификация активных и пассивных действий дает идеальное качество: 1,00;
— классификация направлений движения лифта также показывает показывает очень хорошее качество: 0,99;



(а) (б)

Рис. 8. Зависимость среднего качества от логарифма величины C для активных (а) и пассивных (б) классов. Каждая точка на графике — среднее качество для конкретного значения C на «локальном» скользящем контроле, в рамках одной итерации «глобального». Точки, соответствующие одной итерации «глобального» контроля, объединены в одноцветные ломаные

— одна их трех компонент основной модели — логистическая регрессия с L_1-регуляризацией — использует настройку гиперпараметра C. Этот гиперпараметр настраивается на «локальном» скользящем контроле по схеме, описанной выше, но уже с 13-ю фолдами. Таким образом, получается, своего рода, «вложенный» скользящий контроль. Это необходимо потому, что гиперпараметр C в данной задаче существенно влияет на качество, и его необходимо настроить, и в то же время не переобучиться.
Результаты скользящего контроля приведены на графиках на рис. 8.
Видно, впрочем, что регуляризация для группы пассивных классов дает слабый выигрыш, поскольку большие значения C (и, соответственно, малые значения коэффициента регуляризации C^{-1}) обеспечивают почти оптимальное качество;
— смесь из трех подмоделей (логистическая регрессия, random forest и NN) часто показывает лучшее качество, чем каждая подмодель в отдельности. Это видно по графикам на рис. 9, отображающим качество классификации для разных итераций скользящего контроля, соответствующих отдельным людям;
— **вся композиция** на трех независимых запусках показала качество 0,9202, 0,9216 и 0,9227. Полные результаты классификации приведены в табл. 1.
Как было сказано выше, решающее правило действует по принципу голосования (1). Этот принцип имеет недостаток: могут возникать неопределенности, когда наибольшее количество голосов набирают 2 или более варианта. Однако в данном эксперименте это произошло всего в 9 случаях из 840.
— автору известен только один опубликованный результат классификации набора данных USC-HAD [10] (он также был указан автору рецензентом), однако сравнение с ним не представляется возможным в силу отсутствия описания деталей эксперимента и предложенного метода. Кроме того, насколько можно судить из текста работы, в ходе эксперимента окна, соответствующие одному и тому же замеру, и даже перекрываю-

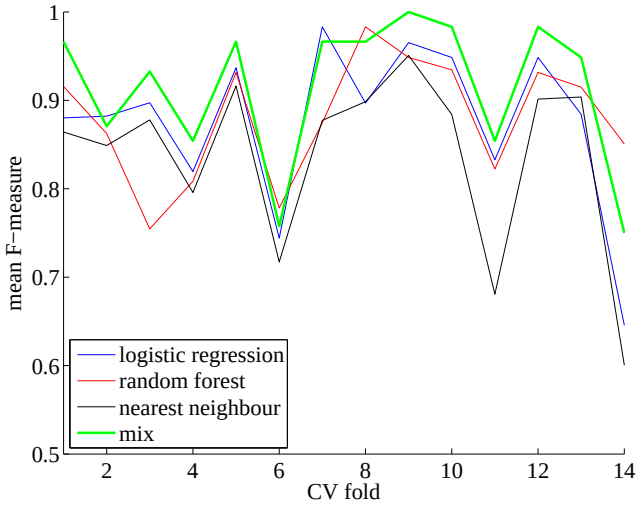


Рис. 9. Качество подмоделей и их смеси на разных итерациях скользящего контроля

Таблица 1. Результаты классификации (confusion matrix). Столбцы соответствуют меткам классов, строки — прогнозам. В каждой ячейке — абсолютное количество соответствующих объектов

	1	2	3	4	5	6	7	8	9	10	11	12
1	70	1	2	6	8	1	0	0	0	0	0	0
2	0	64	2	0	0	0	0	0	0	0	0	0
3	0	2	65	0	0	0	0	0	0	0	0	0
4	0	0	1	64	1	0	1	0	0	0	0	0
5	0	3	0	0	57	1	5	0	0	0	0	0
6	0	0	0	0	3	68	1	0	0	0	0	0
7	0	0	0	0	1	0	63	0	0	0	0	0
8	0	0	0	0	0	0	0	62	8	0	0	0
9	0	0	0	0	0	0	0	3	59	0	2	3
10	0	0	0	0	0	0	0	4	0	70	0	0
11	0	0	0	0	0	0	0	0	2	0	67	1
12	0	0	0	0	0	0	0	1	1	0	1	66
F-мера	0,89	0,94	0,95	0,93	0,84	0,96	0,94	0,89	0,86	0,97	0,96	0,95

щиеся окна могли попасть в обучающую и контрольную выборки, что, как правило, значительно поднимает качество классификации.

Поэтому для сравнения выбран простой метод — логистическая регрессия с L_2-регуляризацией и добавлением константного признака, работающая по принципу «один против всех» на всех 12 классах и всех признаках (всем амплитудном спектре). Иерархия классов никак не используется, но принцип выбора окон и голосования используется тот же.

Такая логистическая регрессия также имеет единственный гиперпараметр C, равный обратному значению коэффициента регуляризации. Метод протестирован всех для значений $C = 2^i$, $i = 0..16$. Результаты приведены на рис. 10. Максимальное качество не превысило значения 0.72.

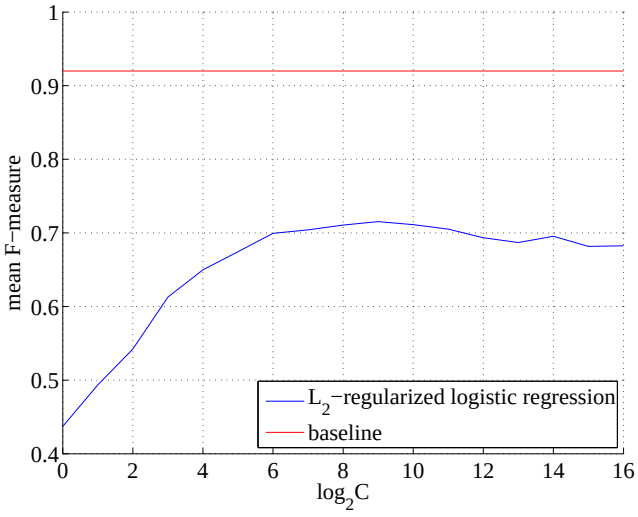


Рис. 10. Качество классификации логистической регрессии с L_2-регуляризацией для различных значений параметра C в сравнении с качеством предложенного метода (baseline)

Заключение

С использованием достаточно простых признаков и относительно сложного по структуре классификатора получено достаточно высокое качество классификации.

Метод получился вполне устойчивым: при небольших изменениях параметров (таких как доля отбираемых признаков, степень перекрытия соседних окон, значения C и т. д.) качество остается в пределах от 0,91 до 0,93. Это можно объяснить тем, что в основной модели используется смесь трех принципиально разных подходов: линейного классификатора, бэггинга над решающими деревьями и метрического классификатора. Эти подмодели по-разному реагируют на изменение параметров, поэтому усредненный результат остается почти неизменным. Остальные же классификаторы композиции очень просты по своей природе и имеют простую интерпретацию.

Наконец, отдельного внимания заслуживает вопрос о применимости предложенного метода на практике. Тут можно отметить следующее:

— предложенный классификатор направления движения лифтов вряд ли представляет практический интерес, но в условиях поставленной задачи это приемлемое решение. Автор постарался лишь удовлетворить формальному требованию классификации, чтобы оценить качество всей композиции. Задача классификации сигналов с единичными «всплесками» несколько выходит за рамки данной работы;
— метод требует достаточно трудоемких вычислений. Однако, во-первых, мощность процессоров постоянно растет, а во-вторых, логистическая регрессия – самая «легковесная» подмодель – очень эффективна с точки зрения вычислительных ресурсов и показывает неплохое качество: 0,89.

Работа выполнена в рамках спецсеминара «Алгебра над алгоритмами и эвристический поиск закономерностей» кафедры Математических методов прогнозирования факультета ВМК МГУ под научным руководством д.ф.-м.н., профессора Дьяконова Александра Геннадьевича. Автор также выражает благодарность Илье Владимировичу Сафонову (Nokia Research) за ценные советы и замечания.

Литература

[1] *Hoseini-Tabatabaei S. A., Gluhak A., Tafazolli R.* A survey on smartphone-based systems for opportunistic user context recognition // *ACM Computing Surveys (CSUR)*, 2013. Vol. 45, No. 3. P. 27.

[2] *Avci A., Bosch S., Marin-Perianu M., Marin-Perianu R., Havinga P.* Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey // *23rd Conference (International) on Architecture of Computing Systems (ARCS)*, 2010. P. 1–10.

[3] *Zhao Z., Chen Y., Liu J., Shen Z., Liu M.* Cross-people mobile-phone based activity recognition // *22nd Joint Conference (International) on Artificial Intelligence Proceedings*. AAAI Press, 2011. Vol. 3. P. 2545–2550.

[4] *Dernbach S., Das B., Krishnan N. C., Thomas B. L., Cook D. J.* Simple and complex activity recognition through smart phones // *8th IEEE Conference (International) on Intelligent Environments*, 2012. P. 214–221.

[5] *Yan Z., Subbaraju V., Chakraborty D., Misra A., Aberer K.* Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach // *16th IEEE Symposium (International) on Wearable Computers (ISWC)*, 2012. P. 17–24.

[6] *Siirtola P., Roning J.* Recognizing Human Activities User-independently on Smartphones Based on Accelerometer Data // *Int. J. Interactive Multimedia Artificial Intelligence*, 2012. Vol. 1, No. 5.

[7] *Incel O. D., Kose M., Ersoy C.* A review and taxonomy of activity recognition on mobile phones // *BioNanoScience*, 2013. Vol. 3, No. 2. P. 145–171.

[8] *Zhang M., Sawchuk A. A.* USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors // *ACM Conference (International) on Ubiquitous Computing (UbiComp) Workshop on Situation, Activity and Goal Awareness (SAGAware)*. Pittsburgh, Pennsylvania, USA, 2012.

[9] *Fan R. E., Chang K W., Hsieh C J, et al.* LIBLINEAR: A library for large linear classification // *J. Machine Learning Res.*, 2008. Vol. 9. P. 1871–1874.

[10] *Cui J., Xu B.* Cost-effective activity recognition on mobile devices // *8th Conference (International) on Body Area Networks Proceedings*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013. P. 90–96.

Поиск эффективных методов снижения размерности при решении задач многоклассовой классификации путем её сведения к решению бинарных задач*

М. Е. Карасиков[1]***, Ю. В. Максимов***[1,2]
karasikov@phystech.edu
[1]МФТИ; [2]ИППИ РАН

Работа посвящена задаче многоклассовой классификации высокой размерности. Рассмотрены способы решения задачи многоклассовой классификации на основе сведения её к задачам бинарной классификации. Исследованы различные подходы к сведению задачи многоклассовой классификации к задачам бинарной классификации и проведено сравнение их эффективностей. Предложены пути повышения производительности классификаторов путем снижения размерности пространства признаков методом случайных проекций. Проведены эксперименты на реальных данных для различных классификаторов, результаты которых отражают характерные зависимости качества классификации и сложности обучения при снижении размерности методом случайных проекций.

Ключевые слова: многоклассовая классификация; One-vs-All; One-vs-One; Error-Correcting Output Codes; лемма Джонсона-Линденштраусса; снижение размерности; случайные проекции.

Dimensionality reduction for multi-class learning problems reduced to multiple binary problems*

M. E. Karasikov[1]***, Y. V. Maximov***[1,2]
[1]MIPT; [2]IITP RAS

Modern machine learning problems, such as image classification, video recognition, text retrieval or engineering diagnostics, leads to the analysis of multi-class learning methods for high-dimensional datasets which can not be solved without data pre-processing. Principal Component Analysis and its randomized versions are some of the most widespread dimensionality reduction methods. We analyze the classification performance of various approaches to multi-class classification (One-vs-One, One-vs-All, Error-Correcting Output Codes) in combination with the dimensionality reduction based on Random Gaussian Projections. Computational efficiency of the Random Projections distinguishes it from other dimensionality reduction methods. With that, low-distortion property of this mapping allows to reduce dimensionality thrice and more with imperceptible quality losses. This leads to an effective and computationally cheap approach for solving multi-class problems in high-dimensional space. Basic theoretical foundations of the approach as well as its computational complexity analysis are discussed. Numerical stability and quality of the method proposed is supported by empirical evaluation of the approach. We provide a number of experiments for different machine learning methods over various real datasets from the open-source machine learning repositories. Experiments show applicability of Random Projections for cheap selection of the most suitable classifier, its parameters optimization and multi-class classification approach selection.

*Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-31241 мол_а.

Keywords: multi-class classification; One-vs-All; One-vs-One; Error-Correcting Output Codes; Johnson-Lindenstrauss lemma; dimensionality reduction; Random Projections.

Введение

В современных задачах машинного обучения часто возникают задачи многоклассовой классификации, в которых множество некоторых объектов нужно отобразить на множество классов большой мощности. Например, распознавание символов по изображению, классификация речи и текста, медицинская диагностика.

Формально задача классификации заключается в следующем. Пусть $X = \{x_1, \ldots, x_\ell\}$ — множество описаний объектов, $Y = \{y_1, \ldots, y_N\}$ — конечное множество меток классов. Существует целевая функция — отображение $y : X \to Y$, значения которого известны только на объектах обучающей выборки

$$\mathfrak{D} = \left\{(x^1, y^1), \ldots, (x^m, y^m)\right\} \subset X \times Y.$$

Требуется построить алгоритм $a : X \to Y$ — отображение, приближающее целевую функцию y на множестве X. Задачу классификации с $N = 2$ ($N > 2$) классами будем называть бинарной (многоклассовой) задачей. В бинарной задаче положим для удобства $Y = \{-1, +1\}$, а в многоклассовой — $Y = \{1, \ldots, N\}$. Для решения многоклассовой задачи можно использовать два способа. Первый способ состоит в использовании многоклассовых классификаторов, например, решающих деревьев. В нашей работе для решения многоклассовой задачи применяется способ, основанный на использовании классификаторов (линейный дискриминант Фишера [1], SVM [2]), решающих бинарные задачи. При этом многоклассовая задача разбивается на множество бинарных задач, которые решаются независимо с использованием техник бинарной классификации. Процесс разбиения будем называть сведением многоклассовой задачи к бинарным.

Общие сведения о задаче многоклассовой классификации даны в [3, 4]. В [4], также, проведено сравнение различных подходов к сведению многоклассовой задачи к бинарным.

Подходы к сведению многоклассовой задачи к бинарным

— **One-vs-All approach** (OVA) заключается в обучении N классификаторов по следующему принципу

$$f_i(x) = \begin{cases} \geqslant 0, & \text{если } y(x) = i, \\ < 0, & \text{если } y(x) \neq i, \end{cases}$$

которые отделяют каждый класс от остальных. Далее, для каждого $x \in X$ вычисляются все классификаторы и выбирается класс, соответствующий классификатору с большим значением:

$$a(x) = \arg\max_{i=1,\ldots,N} f_i(x).$$

— **One-vs-One approach**. Его так же называют All-vs-All (AVA) approach. В этом случае строятся $N(N-1)$ классификаторов, которые разделяют объекты пар различных классов:

$$f_{ij}(x) = \begin{cases} +1, & \text{если } y(x) = i, \\ -1, & \text{если } y(x) = j. \end{cases}$$

После обучения бинарных классификаторов решение принимается следующим образом:

$$a(x) = \arg\max_{i=1,\ldots,N} \sum_{\substack{j=1,\ldots,N \\ j\neq i}} f_{ij}(x).$$

Сравнение первых двух подходов проведено в [5]. Недостаток подхода OVA состоит во многих случаях отказа от классификации [5]. Однако на распознанных объектах этот способ дает очень хорошие результаты [5].

— **Error-Correcting Output Codes approach** (ECOC) предложен в [6]. ECOC предполагает кодирование меток классов двоичными числами длины F, которое сводит задачу определения неизвестного класса объекта x к определению F неизвестных бит кодового слова класса $y(x)$. Для каждого бита строится бинарный классификатор, отделяющий группу классов со значением $+1$ соответствующего бита от классов со значением -1. Пусть $\mathbf{M} \in \{-1, +1\}^{N\times F}$ — кодовая матрица, в строках которой записаны коды меток классов из Y. Тогда обучаются F классификаторов $f_1, \ldots, f_F$ так, чтобы $f_j(x) = i$ тогда и только тогда, когда $M^i_j = 1$. При классификации нового объекта x вычисляется его кодовое слово $\boldsymbol{f}(x) = [f_1(x), \ldots, f_F(x)]$ и выбирается класс, с ближайшим к $\boldsymbol{f}(x)$ кодовым словом. Для расстояния Хэмминга получим:

$$a(x) = \arg\min_{i=1,\ldots,N} \sum_{j=1}^{F} \left(\frac{1 - \operatorname{sign}(M^i_j f_j(x))}{2} \right).$$

В работе [7] было представлено улучшение ECOC, согласно которому кодовая матрица M допускает нулевые элементы, а классификация происходит по правилу

$$a(x) = \arg\min_{i=1,\ldots,N} \sum_{j=1}^{F} L(M^i_j f_j(x)),$$

где L — некоторая функция потерь. В результате наблюдалось снижение числа ошибок классификации почти на всех представленных тестах.
Реализация изложенных подходов представлена в [8].

Как видно [3, 4, 5, 6, 7], существует множество подходов к сведению многоклассовой задачи к бинарным.

Обратимся теперь к вопросу обучения бинарных классификаторов. Как правило, объекты $x \in X$ задаются векторами в n-мерном евклидовом пространстве $\mathbb{R}^n$, которое мы будем называть пространством признаков. Тогда в качестве бинарных классификаторов, как упоминалось выше, можно использовать линейный дискриминант Фишера [1], SVM [2], а так же строить такие композиции, как AdaBoost [9, 10]. Временная сложность обучения и тестирования названных классификаторов как минимум линейна по числу признаков, и в задачах высокой размерности, т. е. с высокоразмерным пространством признаков (например, в задачах распознавания лиц) все эти методы обладают недостаточно высоким быстродействием. Кроме того, из-за линейной сложности по памяти задачи классификации сверх большой размерности решать без предобработки данных не удается. Таким образом, снижение размерности задачи многоклассовой классификации представляется актуальной задачей.

Часто бывает, что в задаче классификации высокой размерности множество объектов $X \subset \mathbb{R}^n$ лежит на линейном многообразии, размерность которого много меньше размерности исходного пространства признаков. В таких случаях чрезвычайно эффективным оказывается метод главных компонент [11]. Метод главных компонент находит ортогональные направления $\mathbf{w}_1, \ldots, \mathbf{w}_d$ (главные компоненты), вдоль которых выборочная дисперсия максимальна, и проецирует элементы множества X на линейное многообразие

$$\bar{\mathbf{x}} + \mathrm{Lin}(\mathbf{w}_1, \ldots, \mathbf{w}_d).$$

Однако метод главных компонент применим не к любым данным, ведь легко привести пример задачи классификации, для которой классификация значительно усложнится после его применения. Таким образом, одна из проблем метода главных компонент — это зависимость от данных. Еще одним недостатком метода главных компонент является высокая временная сложность $O(\ell n^2 + n^3)$ [12].

Альтернативным методом снижения размерности является метод случайных проекций, заключающийся в случайном проецировании исходного пространства признаков на пространство меньшей размерности. Этот метод будет подробно изложен в главе 10. Сложность генерации проекционной матрицы $O(nd)$, где d — размерность редуцированного пространства признаков. Временная сложность нахождения нового признакового описания объектов — $O(\ell dn)$. Преимуществами последнего метода являются его простота и независимость от исходных данных. При всем этом, метод случайных проекций сравним с методом главных компонент по качеству классификации в редуцированном пространстве [13].

В нашей работе анализируется метод случайных проекций как метод снижения размерности задачи многоклассовой классификации. Приводится сравнение эффективностей основных подходов к сведению многоклассовой задачи к бинарным (One-vs-One, One-vs-All, ECOC) при использовании метода случайных проекций.

Постановка задачи

Дана задача многоклассовой классификации с пространством признаков высокой размерности $X \subset \mathbb{R}^n$. Применяя отображение $A_d : X \to X' \subset \mathbb{R}^d,\ d < n$, будем снижать размерность задачи, тем самым понижая сложность классифицирующего алгоритма. При этом может исказиться метрика, то есть изменятся относительные расстояния между объектами в X', что может привести к потере качества классификации. Чем больше сжатие $\frac{n}{d}$, тем больше возможны потери в качестве классификации. Будем искать то минимальное значение d, при котором ошибка классификации e, определенная некоторым функционалом качества, не превышает некоторой заданной допустимой ошибки e'.

Поставим задачу формально. $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\} \subset \mathbb{R}^n$ — множество описаний объектов в евклидовом пространстве, $Y = \{1, \ldots, N\},\ N > 2$ — множество меток классов. Предполагается существование целевой зависимости — отображения $y : X \to Y$, значения которого известны только на объектах обучающей выборки

$$\mathfrak{D} = \left\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^m, y^m)\right\} \subset X \times Y.$$

Пусть задан метод $\mu : \tilde{\mathfrak{D}} \mapsto a$, который $\forall k \geqslant 1$ по произвольной выборке

$$\tilde{\mathfrak{D}} = \left\{(\tilde{\mathbf{x}}^1, y^1), \ldots, (\tilde{\mathbf{x}}^m, y^m)\right\} \subset \mathbb{R}^k \times Y$$

строит алгоритм классификации $a : X \to Y$, решающий задачу многоклассовой классификации с ошибкой $e(a)$, и задана допустимая ошибка классификации e'.

Найти

$$d^* = \min_{e(\mu(A_d(\mathfrak{D})))\leqslant e'} d,$$

где $e(\mu(A_d(\mathfrak{D})))$ — ошибка алгоритма классификации, построенного методом μ по выборке $A_d(\mathfrak{D})$ — образу снижающего размерность отображения A_d, e' — допустимая ошибка классификации.

Многоклассовая классификация

Бинарные классификаторы. Для бинарной классификации, где $Y = \{-1, +1\}$, используются линейные классификаторы $a(\mathbf{x}, \boldsymbol{w}) = \operatorname{sign} f_{\boldsymbol{w}}(\mathbf{x})$, где $f_{\boldsymbol{w}}(\mathbf{x})$ — дискриминантная функция, $\boldsymbol{w}$ — вектор параметров. Обучение классификатора производится путем минимизации эмпирического риска

$$Q(f_{\boldsymbol{w}}, \mathfrak{D}) = \frac{1}{m}\sum_{i=1}^{m} \mathscr{L}\left(y^i f_{\boldsymbol{w}}(\mathbf{x}^i)\right), \tag{1}$$

где функция потерь $\mathscr{L}$ — невозрастающая и неотрицательная. При этом $y^i f_{\boldsymbol{w}}(\mathbf{x}^i)$ называется отступом объекта $\mathbf{x}^i$ относительно алгоритма классификации

$$a(\mathbf{x}, \boldsymbol{w}) = \operatorname{sign} f_{\boldsymbol{w}}(\mathbf{x}). \tag{2}$$

В нашей работе в качестве линейных классификаторов берутся различные модификации SVM и AdaBoost.

В **SVM** [2] за функцию потерь принимается кусочно-линейная функция

$$\mathscr{L}\left(y^i f_{\boldsymbol{w}}(\mathbf{x}^i)\right) = (1 - y^i f_{\boldsymbol{w}}(\mathbf{x}^i))_+,$$

где $(\cdot)_+ = \max\{\cdot, 0\}$. Классификатор $f_{\boldsymbol{w}}$ ищется в виде $f_{\boldsymbol{w}}(\mathbf{x}) = \boldsymbol{w} \cdot \mathbf{x}$, где $\boldsymbol{w}$ — решение задачи безусловной минимизации

$$\frac{1}{2C}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m}(1 - y^i(\boldsymbol{w} \cdot \mathbf{x}^i))_+ \to \min_{\boldsymbol{w}\in\mathbb{R}^n}.$$

Согласно алгоритму **AdaBoost** [9, 10] дискриминантная функция $f_T(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$ конструируется из T базовых классификаторов $\{h_t(\mathbf{x})\}_{t=1}^{T} \subset H$, которые обучаются последовательно так, чтобы минимизировать эмпирический риск $Q(f_T, \mathfrak{D})$ (см. формулу 1) с экспоненциальной функцией потерь $\mathscr{L}(M) = e^{-M}$.

В нашей работе базовые классификаторы выбираются из множества

$$H = \left\{h^{\pm}_{j\theta}(\mathbf{x}) = \pm\operatorname{sign}(x_j - \theta) :\ \mathbf{x} = (x_1, \ldots, x_n),\ j = 1, \ldots, n,\ \theta \in \mathbb{R}\right\}. \tag{3}$$

Подходы к сведению многоклассовой задачи к бинарным. Как было сказано выше, в работе предлагается решать задачу многоклассовой классификации путем сведения её к бинарным задачам. При этом удобно использовать конструкцию, предложенную

в [7]. Эта конструкция предполагает кодирование меток классов $i \in Y = \{1, \ldots, N\}$ строками $M^i \in \{-1, 0, +1\}^F$ длины F, составляющими кодовую матрицу $[M^i_j]^{N \times F}$. При этом для каждого столбца M_j, $j = 1, \ldots, F$, матрицы M получаем бинарную задачу, которая заключается в разделении классов

$$Y_j^{+1} = \{i \in Y : \ M_j^i = +1\} \ \text{ и } \ Y_j^{-1} = \{i \in Y : \ M_j^i = -1\}.$$

Формально каждая бинарная задача выглядит следующим образом. $X_j = X$ — множество объектов, $Y_j = \{-1, +1\}$ — множество меток классов. Построить классификатор $a_j : X_j \to Y_j$, аппроксимирующий целевую функцию $y_j : X_j \to Y_j$, значения которой известны только на объектах обучающей выборки

$$\mathfrak{D}_j = \left\{(\mathbf{x}, M_j^y) : \ M_j^y \neq 0, \ (\mathbf{x}, y) \in \mathfrak{D}\right\},$$

где $M \in \{-1, 0, +1\}^{N \times F}$ — матрица, строки которой состоят из кодов меток классов Y.

На этапе распознавания для каждого нового объекта $\mathbf{x}$ вычисляются все F классификаторов $a_1(\mathbf{x}), \ldots, a_F(\mathbf{x})$, и классификация происходит голосованием, то есть объект $\mathbf{x}$ относится к тому классу, который чаще всего встречается в множествах $Y_1^{a_1(\mathbf{x})}, \ldots, Y_F^{a_F(\mathbf{x})}$:

$$a(\mathbf{x}) = \arg\max_{i=1,\ldots,N} \sum_{j=1}^{F} \mathbf{1}\left[i \in Y_j^{a_j(\mathbf{x})}\right] = \arg\min_{i=1,\ldots,N} \sum_{j=1}^{F} \mathbf{1}[a_j(\mathbf{x}) \neq M_j^i]. \tag{4}$$

Этот метод обобщается выбором произвольного расстояния $d(\cdot, \cdot)$ между кодовыми словами M^i классов $i \in Y$ и словом-ответом $\boldsymbol{a}(\mathbf{x}) = [a_1(\mathbf{x}), \ldots, a_F(\mathbf{x})]$:

$$a(\mathbf{x}) = \arg\min_{i=1,\ldots,N} d(M^i, \boldsymbol{a}(\mathbf{x})).$$

Если каждый классификатор $a_j(\mathbf{x})$ задается дискриминантной функцией $f_j(\mathbf{x})$ (см. 2), то расстояние можно определить произвольной функцией потерь L :

$$d_L(M^i, \boldsymbol{f}(\mathbf{x})) = \sum_{j=1}^{F} L(M_j^i f_j(\mathbf{x})).$$

Тогда классификация будет происходить следующим образом:

$$a(\mathbf{x}) = \arg\min_{i=1,\ldots,N} d_L(M^i, \boldsymbol{f}(\mathbf{x})), \quad \boldsymbol{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_F(\mathbf{x})]. \tag{5}$$

Рассмотрим подробно построение кодовой матрицы M. Для подхода One-vs-All число бинарных классификаторов равно числу классов: $F = N$, а каждый бинарный классификатор $a_j(\mathbf{x})$ $(j = 1, \ldots, F)$ обучается так, чтобы отделять объекты класса с меткой j от остальных объектов множества X. Тогда метки классов кодируются следующим образом:

$$M = \begin{pmatrix} +1 & -1 & \ldots & -1 \\ -1 & +1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & +1 \end{pmatrix}_{N \times N}.$$

Для подхода All-vs-All число классификаторов равно $F = \binom{N}{2}$, и классификаторы разделяют каждую пару классов. Для этого подхода кодовая матрица M записывается в виде:

$$M = \begin{pmatrix} +1 & +1 & \dots & +1 & +1 & 0 & \dots & \dots & 0 \\ -1 & 0 & \dots & 0 & 0 & +1 & \dots & \dots & 0 \\ 0 & -1 & \dots & 0 & 0 & -1 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & -1 & 0 & 0 & \dots & \dots & +1 \\ 0 & 0 & \dots & 0 & -1 & 0 & \dots & \dots & -1 \end{pmatrix}_{N \times \binom{N}{2}}.$$

Подход ECOC [6] обобщает подходы One-vs-All и All-vs-All. Его преимуществом является возможность использования кодов, исправляющих ошибки. Пусть минимальное расстояние Хэмминга между строками кодовой матрицы M равно $d_{\min}$. Тогда корректирующая способность кода равна $t = \left[\frac{d_M - 1}{2}\right]$. Свойство кода исправлять ошибки повышает точность классификации, т. к. код гарантированно восстанавливает исходное слово, если произошло не более t ошибок бинарных классификаторов. В этом случае многоклассовая классификация (см. формулу 4) происходит корректно. Таким образом, высокая корректирующая способность кода уменьшает число ошибок многоклассовой классификации. Однако с ростом длины кода вместе с корректирующей способностью растет и число бинарных классификаторов, а, значит, и среднее число ошибок. Так как корректирующая способность t такого кода растет линейно вместе с длиной кода, как и число ошибок бинарных классификаторов, то улучшение качества многоклассовой классификации за счет увеличения длины кода ограничено. В нашей работе для построения кодовой матрицы M в подходе ECOC применяются БЧХ коды и случайное кодирование. При случайном кодировании генерируется множество случайных кодовых матриц и из них выбирается матрица с максимальной корректирующей способностью.

Снижение размерности

В последующих рассуждениях этого раздела будем опираться на работу [14]. Для снижения размерности пространства признаков будем использовать случайное линейное отображение

$$A: \quad \mathbb{R}^n \supset X \to X' \subset \mathbb{R}^d. \tag{6}$$

Пусть $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_\ell]$ — транспонированная матрица объектов-признаков в исходном пространстве признаков, $\mathbf{A} = [\xi_{ij}]^{d \times n}$ — проекционная матрица, соответствующая отображению 6, где ξ_{ij} — независимые центрированные одинаково распределенные случайные величины с мат. ожиданием $\mathsf{E}\xi_{ij} = 0$ и дисперсией $\mathsf{D}\xi_{ij} = \sigma^2$. Транспонированная матрица объектов-признаков в редуцированном пространстве запишется в виде $\mathbf{X}' = \mathbf{A}\mathbf{X}$.

Будем искать отображения 6, сохраняющие попарные расстояния между объектами множества X с точностью $\varepsilon \in (0, 1)$, аналогично условию леммы Джонсона-Линденштрауса [14]:

$$\forall \mathbf{x}', \mathbf{x}'' \in X \quad (1 - \varepsilon)\|\mathbf{x}' - \mathbf{x}''\|_2^2 < \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}''\|_2^2 < (1 + \varepsilon)\|\mathbf{x}' - \mathbf{x}''\|_2^2. \tag{7}$$

$$\|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}''\|_2^2 = (\mathbf{x}' - \mathbf{x}'')^\intercal \mathbf{A}^\intercal \mathbf{A}(\mathbf{x}' - \mathbf{x}'') = (\mathbf{x}' - \mathbf{x}'')^\intercal (\mathbf{I} + \underbrace{\mathbf{A}^\intercal \mathbf{A} - \mathbf{I}}_{\mathbf{\Sigma}})(\mathbf{x}' - \mathbf{x}'')$$

Условие 7 можно переписать в виде

$$\forall \mathbf{x}', \mathbf{x}'' \in X \quad -\varepsilon\|\mathbf{x}' - \mathbf{x}''\|_2^2 < (\mathbf{x}' - \mathbf{x}'')^\mathsf{T}\mathbf{\Sigma}(\mathbf{x}' - \mathbf{x}'') < \varepsilon\|\mathbf{x}' - \mathbf{x}''\|_2^2.$$

Логично брать такую проекционную матрицу $\mathbf{A}$, чтобы $\mathsf{E}\mathbf{\Sigma} = 0$.

$$\mathsf{E}[\mathbf{\Sigma}]_{ij} = \mathsf{E}[\mathbf{A}^\mathsf{T}\mathbf{A} - \mathbf{I}]_{ij} = \mathsf{E}\sum_{k=1}^{d} \xi_{ki}\xi_{kj} - \delta_{ij} = (\sigma^2 d - 1)\delta_{ij},$$

где $\delta_{ij} = \mathbf{1}[i = j]$ — символ Кронекера. Положим $\sigma^2 = \frac{1}{d}$.

Оценим вероятность того, что случайная матрица $\mathbf{A} = [\xi_{ij}]^{d\times n}$ независимых гауссовских случайных величин $\xi_{ij} \sim \mathcal{N}(0, \frac{1}{d})$ удовлетворяет условию 7. Для этого сначала оценим вероятность того, что отображение A сильно изменит длину некоторого фиксированного вектора $\mathbf{u} \in \mathbb{R}^n$. Пусть ортогональная матрица

$$\mathbf{C}: \quad \mathbf{C}\mathbf{u} = [\underbrace{\|\mathbf{u}\|_2, 0, \ldots, 0}_{n}]^\mathsf{T} = \|\mathbf{u}\|_2\boldsymbol{e}_1, \quad \mathbf{C}^\mathsf{T}\mathbf{C} = \mathbf{C}\mathbf{C}^\mathsf{T} = \mathbf{I}.$$

$$\|\mathbf{A}\mathbf{u}\|_2^2 = \mathbf{u}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{u} = \mathbf{u}^\mathsf{T}\mathbf{C}^\mathsf{T}\underbrace{\mathbf{C}\mathbf{A}^\mathsf{T}}_{\mathbf{B}^\mathsf{T}}\underbrace{\mathbf{A}\mathbf{C}^\mathsf{T}}_{\mathbf{B}}\mathbf{C}\mathbf{u} = \|\mathbf{B}\mathbf{C}\mathbf{u}\|_2^2 = \|\mathbf{u}\|_2^2\|\mathbf{B}\mathbf{e}_1\|_2^2.$$

Поскольку $\mathbf{A} \sim \mathcal{N}^{d\times n}\left(0, \frac{1}{d}\right)$, то $\mathbf{B} = \mathbf{A}\mathbf{C}^\mathsf{T} \sim \mathcal{N}^{d\times n}\left(0, \frac{1}{d}\right)$.

Тогда получим

$$\begin{aligned}
\mathsf{P}\left\{\left|\|\mathbf{A}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| \geqslant \varepsilon\|\mathbf{u}\|_2^2\right\} &= \mathsf{P}\left\{\left|\|\mathbf{B}\boldsymbol{e}_1\|_2^2 - 1\right| \geqslant \varepsilon\right\} = \\
&= \mathsf{P}\left\{\left|\sum_{i=1}^{d}\xi_{i1}^2 - 1\right| \geqslant \varepsilon\right\} = \\
&= \mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2 \leqslant 1 - \varepsilon\right\} + \mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2 \geqslant 1 + \varepsilon\right\}.
\end{aligned}$$

Оценим вероятности $\mathsf{P}\left\{\sum\limits_{i=1}^{d}\xi_{i1}^2 \leqslant 1 - \varepsilon\right\}$ и $\mathsf{P}\left\{\sum\limits_{i=1}^{d}\xi_{i1}^2 \geqslant 1 + \varepsilon\right\}$ отдельно.

$$\begin{aligned}
\mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2 \leqslant 1 - \varepsilon\right\} &= \\
= \sup_{t<0}\mathsf{P}\left\{\exp\left(t\sum_{i=1}^{d}\xi_{i1}^2\right) \geqslant \exp\left(t\left(1-\varepsilon\right)\right)\right\} &\overset{\text{(неравенство Маркова)}}{\leqslant} \sup_{t<0}\frac{\mathsf{E}\left[\exp\left(t\sum_{i=1}^{d}\xi_{i1}^2\right)\right]}{\exp\left(t\left(1-\varepsilon\right)\right)} = \\
&= \sup_{t<0}\frac{\left(\mathsf{E}\left[\exp\left(t\xi^2\right)\right]\right)^d}{\exp\left(t\left(1-\varepsilon\right)\right)} = \\
&= \sup_{t<0}\left[\left(\int_{-\infty}^{+\infty} e^{tx^2}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}dx\right)^d \exp\left(-t\left(1-\varepsilon\right)\right)\right] = \\
&= \sup_{t<0}\left[\left(\int_{-\infty}^{+\infty} e^{t\frac{x^2}{d}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\right)^d \exp\left(-t\left(1-\varepsilon\right)\right)\right] = \\
&= \sup_{t<0}\left[\left(1-\frac{2t}{d}\right)^{-\frac{d}{2}}\exp\left(-t\left(1-\varepsilon\right)\right)\right].
\end{aligned}$$

$$\frac{d}{dt}\left[\left(1-\frac{2t}{d}\right)^{-\frac{d}{2}}\exp\left(-t\left(1-\varepsilon\right)\right)\right]=0$$
$$\Leftrightarrow 1-\left(1-\frac{2t}{d}\right)\left(1-\varepsilon\right)=0$$
$$\Leftrightarrow t=-\frac{d}{2}\frac{\varepsilon}{1-\varepsilon}.$$

Итак,

$$\mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2\leqslant 1-\varepsilon\right\}\leqslant\sup_{t<0}\left[\left(1-\frac{2t}{d}\right)^{-\frac{d}{2}}\exp\left(-t\left(1-\varepsilon\right)\right)\right]=$$
$$=\left(1+\frac{\varepsilon}{1-\varepsilon}\right)^{-\frac{d}{2}}\exp\left(\frac{d}{2}\varepsilon\right)=\exp\left(\frac{d}{2}\left(\varepsilon+\ln(1-\varepsilon)\right)\right)\leqslant\exp\left(-\frac{d}{2}\frac{\varepsilon^2}{2}\right).$$

Аналогично доказывается

$$\mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2\geqslant(1+\varepsilon)\right\}\leqslant\exp\left(-\frac{d}{2}\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)\right).$$

Возьмем

$$d(\gamma,\varepsilon)=\left\lceil\frac{4\gamma\ln(\ell)}{\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}}\right\rceil,\ \gamma\geqslant 1.$$

Получим оценку для вероятности сильного изменения длины вектора $\mathbf{u}\in\mathbb{R}^n$:

$$\mathsf{P}\left\{\left|\|\mathbf{A}\mathbf{u}\|_2^2-\|\mathbf{u}\|_2^2\right|>\varepsilon\|\mathbf{u}\|_2^2\right\}\leqslant 2\exp\left(-\frac{d(\gamma,\varepsilon)}{2}\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)\right)\leqslant\frac{2}{\ell^{2\gamma}}.$$

Далее можно получить оценку вероятности того, что матрица $\mathbf{A}$ удовлетворяет условию 7:

$$\mathsf{P}\left\{(1-\varepsilon)\|\mathbf{x}'-\mathbf{x}''\|_2^2<\|\mathbf{A}\mathbf{x}'-\mathbf{A}\mathbf{x}''\|_2^2<(1+\varepsilon)\|\mathbf{x}'-\mathbf{x}''\|_2^2\quad\forall\mathbf{x}',\mathbf{x}''\in X\right\}\geqslant$$
$$\geqslant 1-\binom{\ell}{2}\frac{2}{\ell^{2\gamma}}=1-\frac{\ell-1}{\ell^{2\gamma-1}}=1-\ell^{2-2\gamma}+\ell^{1-2\gamma}.$$

Положив $\gamma=1$, получим, что проекционная матрица

$$\mathbf{A}=[\xi_{ij}]^{d\times n},\ \text{где}\ \xi_{ij}\sim\mathcal{N}\left(0,\frac{1}{d(1,\varepsilon)}\right),\ d(1,\varepsilon)=\left\lceil\frac{4\ln(|X|)}{\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}}\right\rceil,$$

удовлетворяет условию 7 с вероятностью не меньшей

$$p\geqslant\frac{1}{|X|}.\tag{8}$$

Заметим, что для генерации случайной матрицы $\mathbf{A}=[\xi_{ij}]^{d\times n}$ можно использовать и другие распределения случайных величин ξ_{ij} с мат. ожиданием $\mathsf{E}\xi_{ij}=0$ и дисперсией $\mathsf{D}\xi_{ij}=\frac{1}{d}$.

Например, в работе [15] рассматривается случай следующего распределения:

$$\xi_{ij} = \begin{cases} +\sqrt{\frac{s}{d}}, & p = \frac{1}{2s}, \\ 0, & p = 1 - \frac{1}{s}, \\ -\sqrt{\frac{s}{d}}, & p = \frac{1}{2s}. \end{cases}$$

Частный случай для $s = 3$ ранее был предложен в работе [16]. С таким распределением ξ_{ij} матрица $\mathbf{A}$ в среднем имеет разреженность $1 - \frac{1}{s}$, что уменьшает сложность метода случайных проекций в s раз.

Заметим также, что многие современные алгоритмы классификации эффективны при работе с разреженными матрицами, поэтому использование метода случайных проекций с такими классификаторами оправдано лишь в тех случаях, когда матрица объектов-признаков $\mathbf{X}$ не является сильно разреженной. Иначе классификация в задаче высокой размерности с разреженной матрицей объектов-признаков может оказаться вычислительно эффективнее классификации в низкоразмерной задаче, где матрица объектов-признаков не является разреженной, т. к. метод случайных проекций не сохраняет разреженность.

Алгоритм

Предлагается следующий алгоритм многоклассовой классификации:

Input: матрица объектов-признаков $\mathbf{X}^{\ell \times n}$

1. Сгенерировать проекционную матрицу $\mathbf{A} \sim \mathcal{N}^{n \times d}\left(0, \frac{1}{d}\right)$.
2. Найти новые описания объектов: $\mathbf{X}' = \mathbf{X}\mathbf{A}$.
3. Решить задачу многоклассовой классификации с новыми описаниями объектов $\mathbf{X}'$.

В следующей главе представлены эксперименты многоклассовой классификации для различных бинарных классификаторов и подходов к сведению многоклассовой задачи к бинарным.

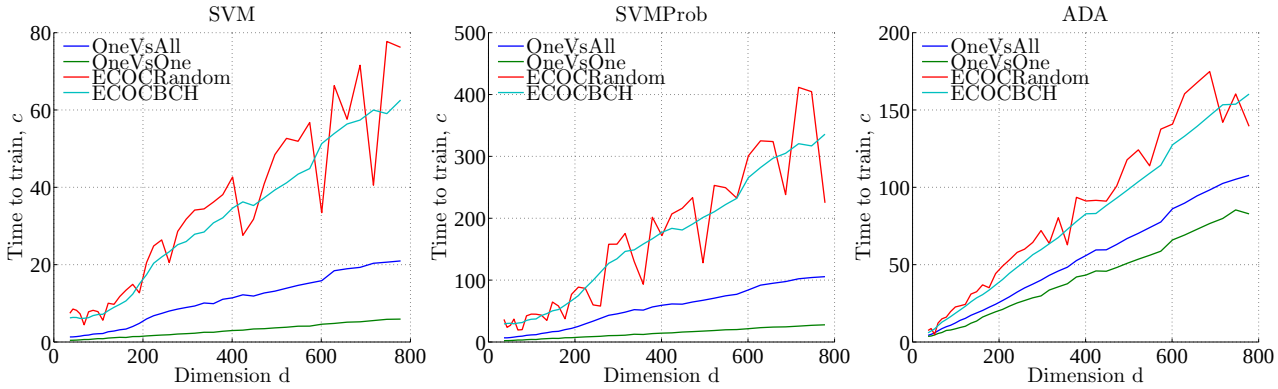
Вычислительный эксперимент

Вычислительный эксперимент проводился с целью демонстрации увеличения производительности бинарных классификаторов SVM и AdaBoost с применением метода случайных проекций как метода снижения размерности задачи многоклассовой классификации. Для сведения исходной многоклассовой задачи к бинарным использовались подходы All-vs-All, One-vs-All и ECOC (BCH и Random). Далее под качеством многоклассовой классификации тестовой выборки X, $|X| = \ell$, будем понимать долю правильных ответов $\frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{1}[a(x_i) = y(x_i)]$. В данном разделе для краткости метод случайных проекций будем называть RP методом.

MNIST dataset

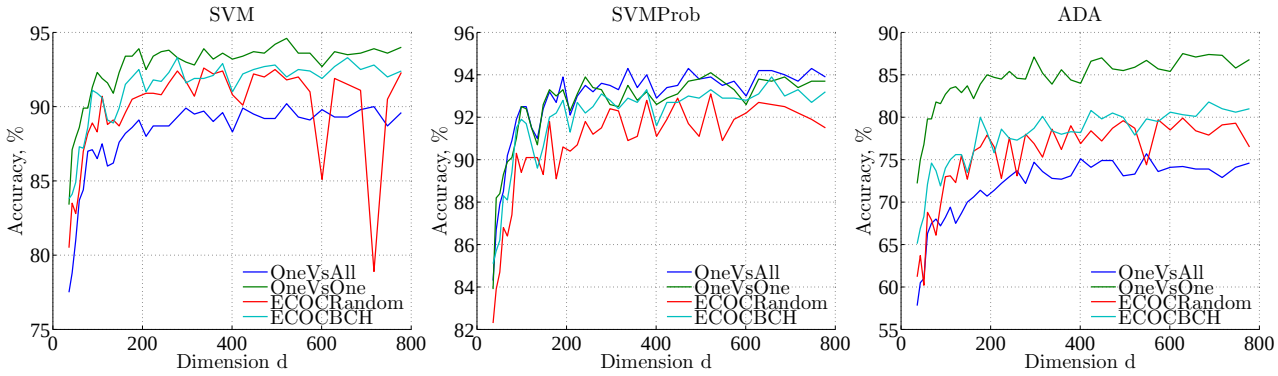
Для вычислительного эксперимента использовалась случайная выборка 2000 объектов для обучения и 1000 объектов для тестирования из базы MNIST [17]:

— of classes: 10
— of data: $60,000$ / $10,000$ (testing)
— of features: 780 / 778 (testing)



(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 1. Зависимость времени обучения бинарных классификаторов Vote SVM, Probabilistic SVM и AdaBoost от размерности редуцированного пространства при одной реализации RP. Data set: MNIST.

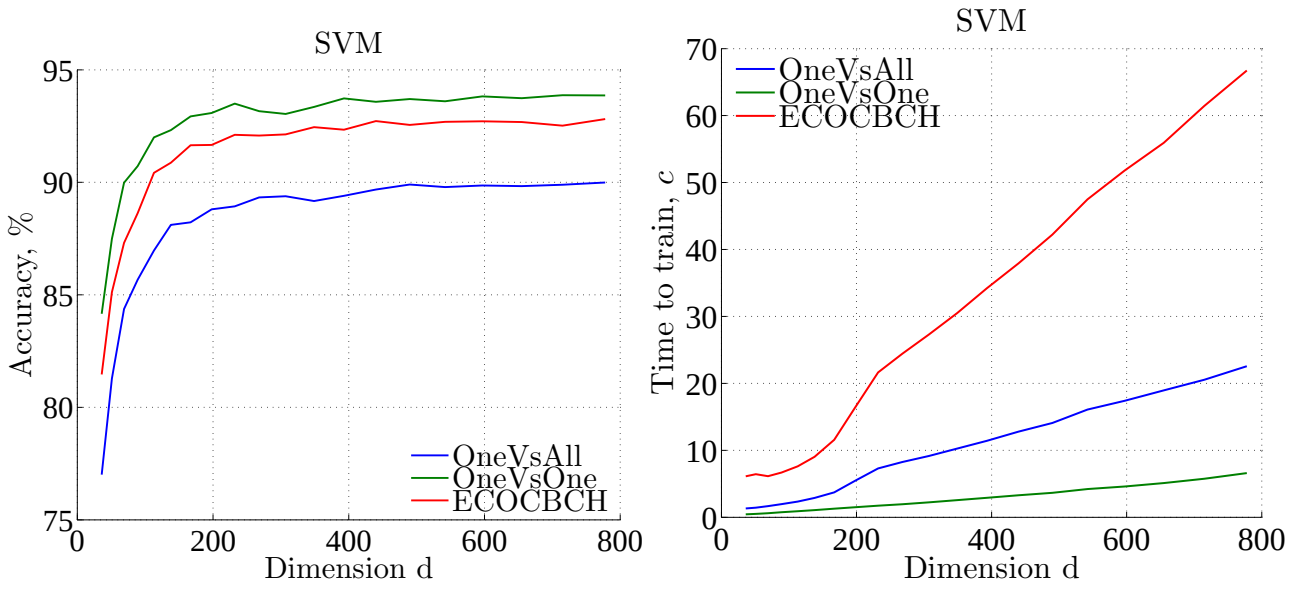


(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 2. Зависимость качества многоклассовой классификации от размерности редуцированного пространства при одной реализации RP. Data set: MNIST.

Задача многоклассовой классификации сводилась к бинарным при помощи подходов OVA, AVA, ECOC-Random (18 столбцов, разреженность 50%) и ECOC-BCH (длина БЧХ кода 15). Задача бинарной классификации решалась классификаторами SVM [18] (Vote SVM), SVMProb [18] (Probabilistic SVM) и AdaBoost [8]. Настройка классификатора SVM производилась с квадратичным ядром и параметрами регуляризации $C = 1$, $\gamma = 1$. В алгоритме AdaBoost использовались 50 базовых классификаторов 3.

Из рисунков 1 и 2 видно, что подход ECOC-BCH оказался предпочтительнее подхода ECOC-Random, так как ECOC-BCH стабильнее по времени обучения и качеству классификации. Можно видеть, что для всех рассмотренных случаев решения задачи многоклассовой классификации зависимость времени обучения от размерности пространства признаков не сильно отличается от линейной. Снизив размерность в 3 раза при помощи RP метода, мы получили тройной прирост в скорости обучения, при этом ошибка возросла на величину порядка $2-3\%$. Рисунок 2 наглядно отображает разброс точности, зависящий от конкретного проецирования.



(а) Binary classifier: Vote SVM (б) Binary classifier: Vote SVM

Рис. 3. Результаты классификации, усредненные по 10 реализациям RP. Data set: MNIST.

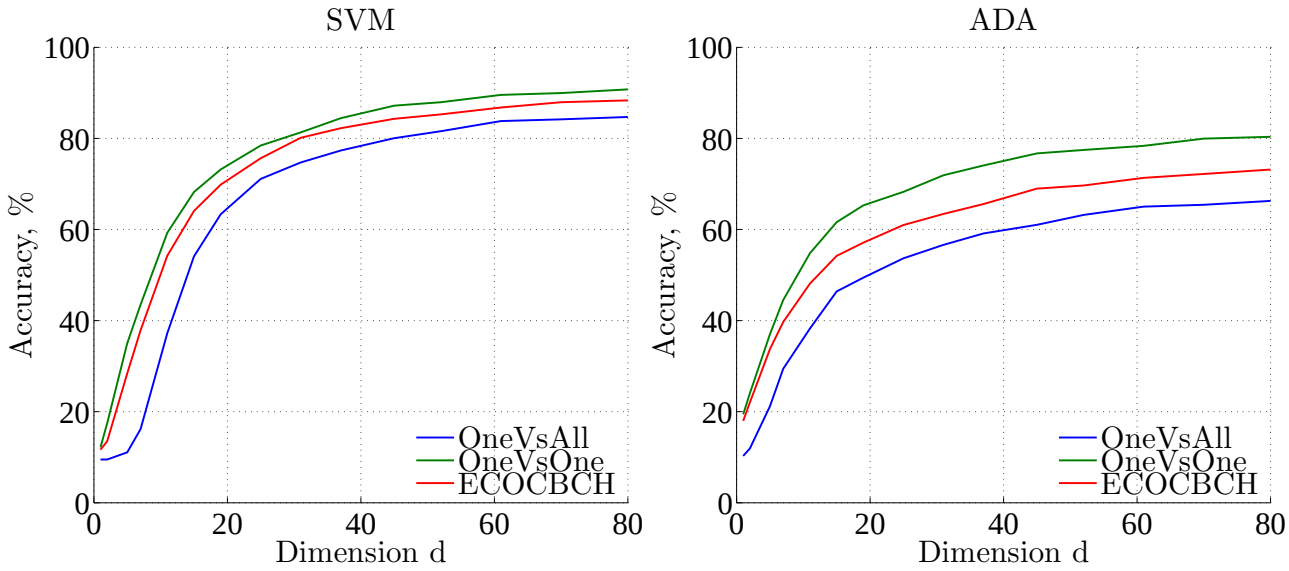
Таблица 1. Качество многоклассовой классификации. Data set: MNIST, Binary classifier: Vote SVM.

Сжатие $\frac{d}{n}$	Подходы											
	One-vs-All			One-vs-One			ECOC-Random			ECOC-BCH		
	min	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*
0.05	76.6	77.3	78.2	82.2	83.3	84.4	75.5	79.6	81.3	81.2	82.1	82.7
0.07	81.2	82.5	85.4	88.3	88.8	89.4	82.9	85.0	86.2	84.6	86.1	87.6
0.11	84.4	85.3	86.4	89.4	90.6	92.5	85.1	87.3	89.9	86.4	88.6	90.1
0.15	86.0	87.0	88.1	91.5	92.1	92.8	85.7	89.2	90.5	89.5	90.1	91.3
0.19	87.5	88.0	88.7	91.1	92.5	93.4	88.8	90.1	91.0	89.7	91.1	92.0
0.25	87.8	88.6	89.2	92.2	92.8	93.5	83.5	89.7	91.3	90.7	91.9	92.8
0.30	88.0	88.8	89.6	92.3	93.2	93.8	89.7	90.6	92.1	90.7	91.8	92.5
0.37	**89.0**	**89.6**	**90.2**	**93.0**	**93.5**	**94.1**	**91.0**	**91.5**	**92.7**	91.3	92.3	93.0
0.44	88.8	89.3	90.1	92.6	93.2	94.0	89.5	**91.3**	92.3	**92.0**	92.3	93.1
0.52	88.9	89.6	90.2	93.1	93.7	94.1	88.8	90.9	92.8	91.8	92.4	**93.8**
0.60	**89.1**	89.8	90.5	93.2	93.7	94.3	**90.5**	**91.5**	92.8	91.4	92.4	93.3
0.69	88.7	89.8	90.5	93.0	**93.8**	94.4	79.4	90.3	92.7	91.9	92.7	93.4
0.79	89.4	89.7	90.4	93.2	**93.8**	94.4	90.4	91.7	92.4	92.3	92.7	**93.7**
0.89	**89.5**	90.0	90.5	**93.4**	93.7	94.1	89.2	91.6	92.7	92.5	92.9	93.5
1.00	89.2	90.0	90.5	93.2	93.7	94.2	91.4	92.2	92.8	92.5	92.9	93.3

На рисунке 3 показаны средние для качества многоклассовой классификации и времени обучения классификаторов Vote SVM.

Детальные результаты с серии 10 реализаций RP представлены в таблице 1.

На рисунке 4 показана усредненная по 20 случайным проекциям зависимость качества классификации для алгоритмов SVM и AdaBoost в сильно редуцированных пространствах

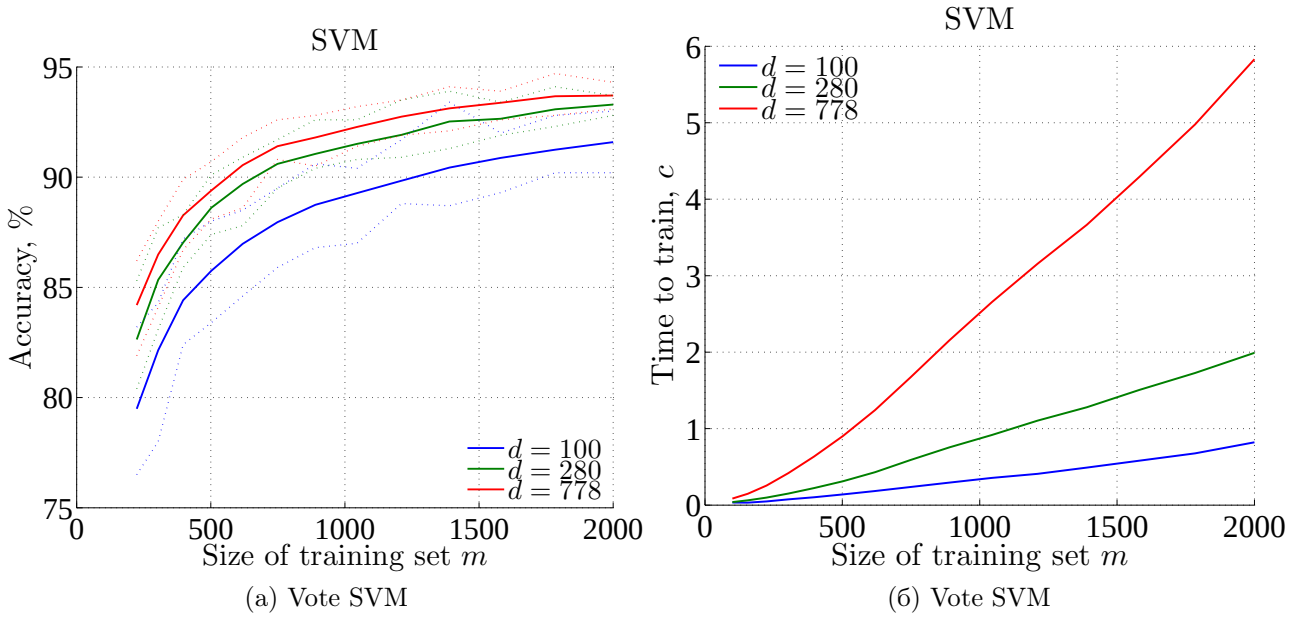


(а) Binary classifier: Vote SVM (б) Binary classifier: AdaBoost

Рис. 4. Качество классификации в сильно редуцированных пространствах, усредненное по 12 реализациям RP. Data set: MNIST.

признаков. На нем видно, что существенные потери в качестве начинаются после сжатия до размерности $d = 30$.

Рисунок 5 показывает устойчивость RP метода относительно мощности обучающей выборки. Тонкими пунктирами обозначены максимальные отклонения от среднего за 20 реализаций RP.



(а) Vote SVM (б) Vote SVM

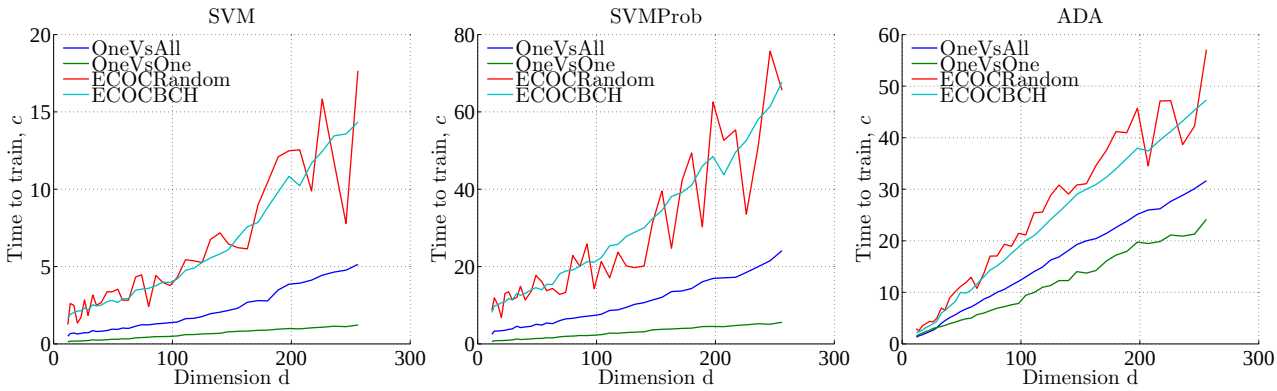
Рис. 5. Зависимость качества многоклассовой классификации от размера обучающей выборки. Усреднение по 20 реализациям RP. Data set: MNIST.

USPS dataset

Использовалась случайная выборка 2000 объектов для обучения и 1000 объектов для тестирования из базы USPS:

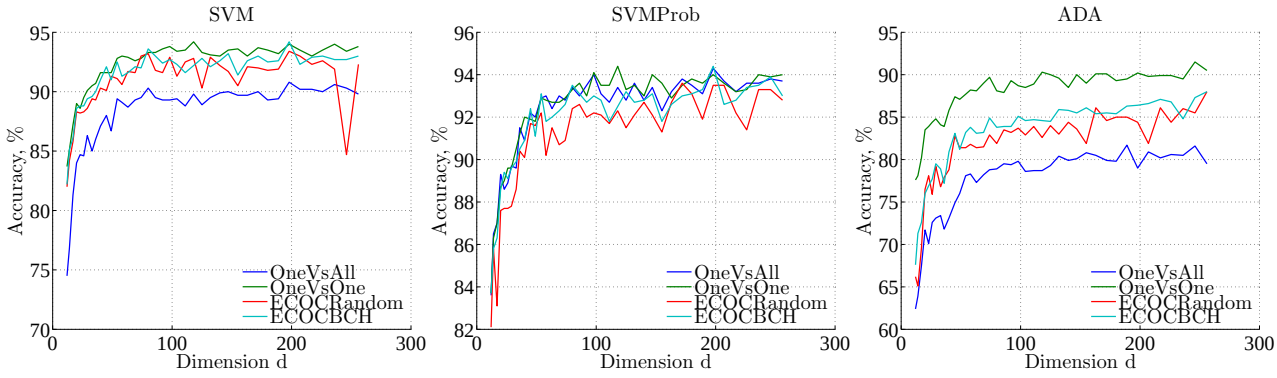
— of classes: 10
— of data: 7,291 / 2,007 (testing)
— of features: 256

В отличие от данных MNIST, матрица объектов-признаков USPS не является разреженной. Многоклассовая задача классификации решалась при тех же условиях, что в эксперименте с данными 10 за исключением того, что ядро бинарного классификатора SVM выбиралось кубическое.



(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 6. Зависимость времени обучения бинарных классификаторов Vote SVM, Probabilistic SVM и AdaBoost от размерности редуцированного пространства при одной реализации RP. Data set: USPS.

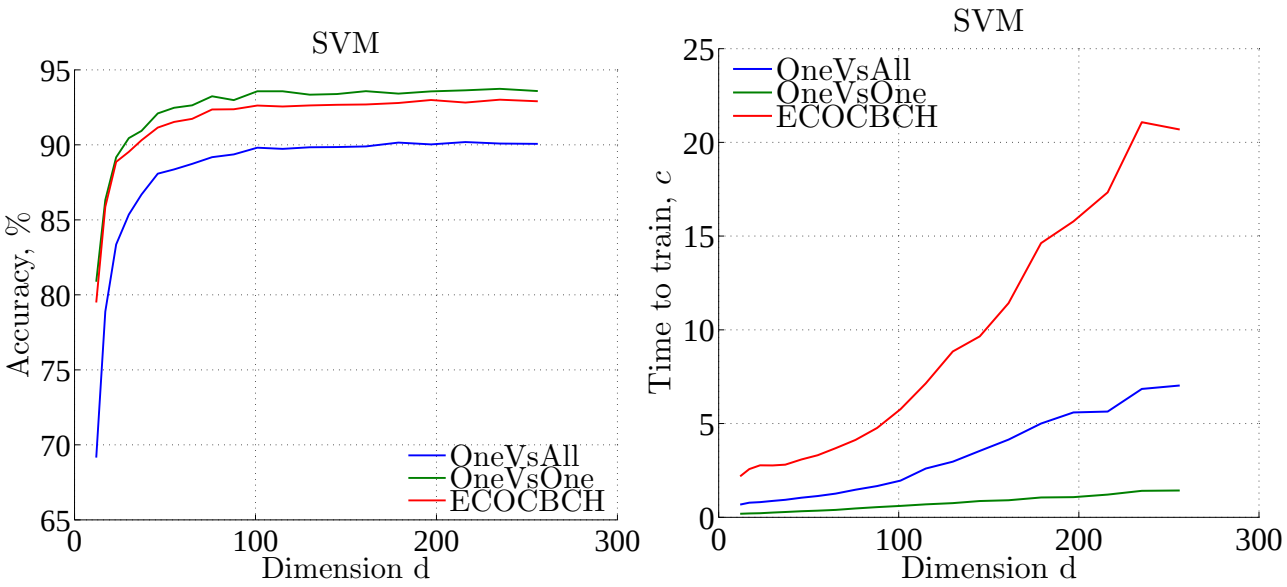


(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 7. Зависимость качества многоклассовой классификации от размерности редуцированного пространства при одной реализации RP. Data set: USPS.

Рисунки 6 и 7, в целом, аналогичны рисункам 1 и 2 из прошлого параграфа.

На рисунке 8 показана зависимость средних для качества многоклассовой классификации и времени обучения классификаторов Vote SVM от размерности редуцированного пространства d.



(а) Binary classifier: Vote SVM (б) Binary classifier: Vote SVM

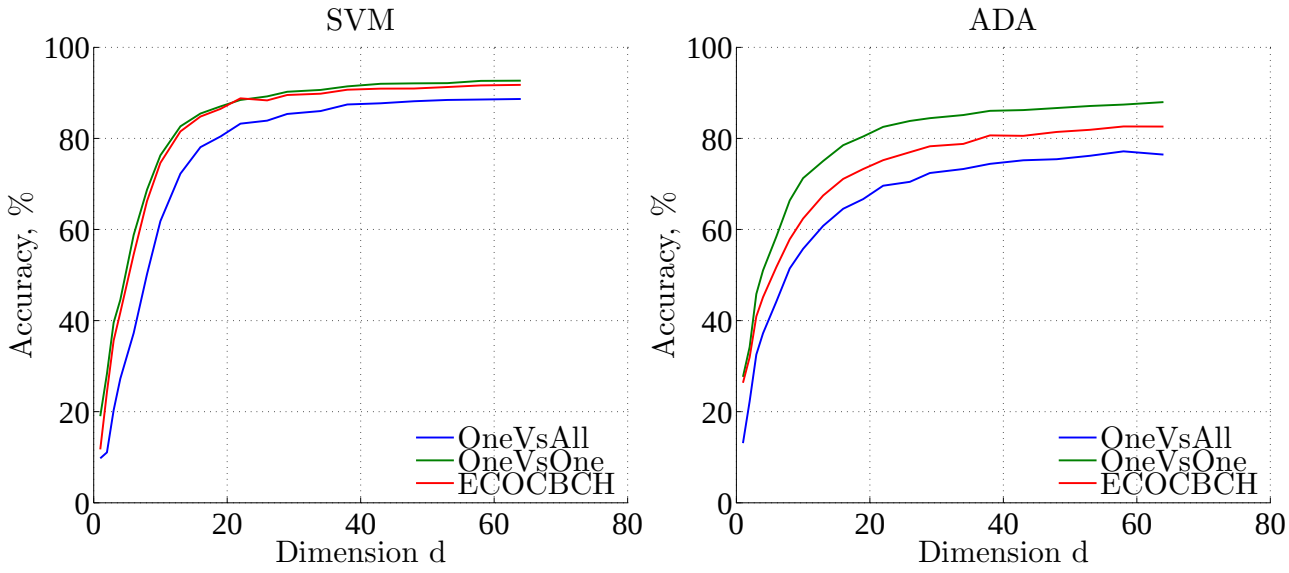
Рис. 8. Результаты классификации, усредненные по 15 реализациям RP. Data set: USPS.

Таблица 2. Качество многоклассовой классификации. Data set: USPS, Binary classifier: Vote SVM.

Сжатие $\frac{d}{n}$	Подходы											
	One-vs-All			One-vs-One			ECOC-Random			ECOC-BCH		
	min	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*
0.05	65.4	69.0	73.8	76.8	80.8	85.6	74.5	78.7	81.0	76.7	79.6	84.7
0.07	78.4	81.4	83.7	84.9	87.1	88.5	83.2	85.5	87.6	85.3	87.0	88.2
0.11	83.7	85.2	87.1	88.9	90.0	91.5	87.1	88.6	90.3	87.9	88.9	90.1
0.15	86.0	87.1	88.7	90.5	91.6	92.5	87.0	89.5	90.8	89.7	90.9	**92.8**
0.20	86.2	87.8	88.9	91.1	92.2	93.1	86.4	90.3	91.9	90.5	91.4	92.6
0.25	87.3	88.6	89.9	91.7	92.6	93.4	**90.1**	91.2	92.5	90.7	91.6	92.5
0.30	88.4	89.2	90.4	92.0	92.9	93.9	85.7	91.1	92.5	**91.1**	92.1	93.0
0.37	88.2	89.3	90.4	92.7	93.2	93.8	**90.9**	91.8	92.9	90.7	92.1	93.0
0.44	88.1	89.6	90.3	92.8	93.4	94.0	**90.3**	91.8	92.7	91.8	92.5	93.1
0.52	**89.2**	89.8	**91.0**	**93.0**	93.6	94.3	86.8	92.2	**93.5**	**92.1**	**92.9**	**93.7**
0.60	88.2	89.7	90.4	92.8	93.5	94.2	**91.0**	92.2	**93.6**	91.8	92.7	**93.6**
0.69	88.7	89.7	90.8	92.8	93.5	94.0	**91.8**	92.3	92.9	91.8	92.6	93.3
0.79	**89.3**	**90.0**	90.8	92.8	93.6	94.4	86.4	91.9	93.1	91.8	92.7	93.3
0.89	88.6	89.9	91.1	93.3	93.7	94.5	**91.6**	**92.4**	93.1	**92.4**	**93.**0	**93.7**
1.00	89.3	90.2	91.2	93.2	93.6	93.9	84.9	91.6	93.3	92.0	92.7	93.3

Детальные результаты с серии 15 реализаций RP представлены в таблице 2.

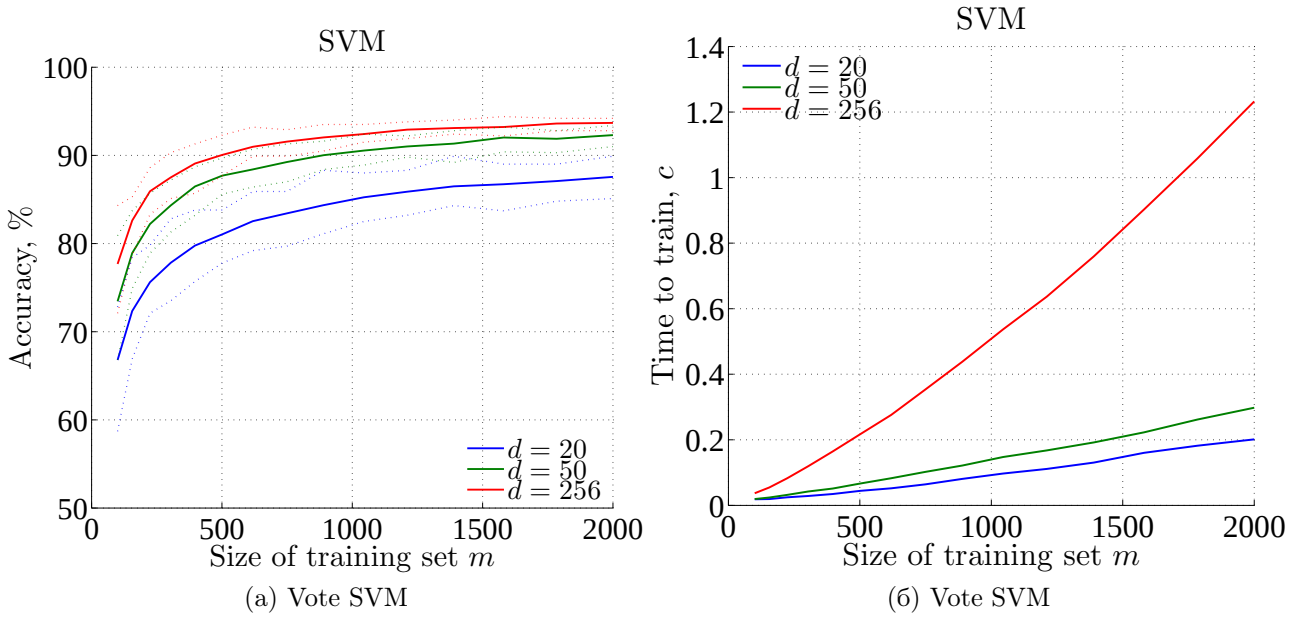
На рисунке 9 показана усредненная по 10 случайным проекциям зависимость качества классификации для алгоритмов SVM и AdaBoost в сильно редуцированных пространствах признаков. На нем видно, что существенные потери в качестве начинаются после сжатия до размерности $d = 20$.



(а) Binary classifier: Vote SVM (б) Binary classifier: AdaBoost

Рис. 9. Качество классификации в сильно редуцированных пространствах, усредненное по 10 реализациям RP. Data set: USPS.

Как и в предыдущем параграфе, рисунок 10 показывает устойчивость RP метода. Тонкими пунктирами обозначены максимальные отклонения от среднего за 40 реализаций RP.



(а) Vote SVM (б) Vote SVM

Рис. 10. Зависимость качества многоклассовой классификации от размера обучающей выборки. Усреднение по 40 реализациям RP. Data set: USPS.

Выводы

В проведенных экспериментах было выяснено, что RP метод снижения размерности наиболее устойчив при использовании подходов One-vs-One и ECOC-BCH. Подход ECOC-Random является наименее устойчивым, что видно на рисунках 2 и 7. Однако

средняя зависимость качества многоклассовой классификации для всех рассмотренных подходов One-vs-All, One-vs-One, ECOC-Random и ECOC-BCH приблизительно одинаковая с точностью до сдвига, продиктованного выбором функции потерь в формуле 5 и геометрией задачи, избирающей приоритетный подход.

Заключение

В работе было показано, что предложенный метод случайных проекций снижения размерности задачи многоклассовой классификации устойчив по отношению к обучающей выборке для всех рассмотренных подходов к сведению многоклассовой задачи классификации к множеству бинарных задач. Метод случайных проекций, как правило, позволяет снизить размерность в два – четыре раза с потерей качества решения многоклассовой задачи порядка 5%. Метод слабо зависит от данных, вычислительно эффективен и достаточно прост в применении, что позволяет использовать его в частности при прототипировании алгоритмов анализа данных.

Результаты экспериментов свидетельствуют о том, что наиболее эффективна работа метода случайных проекций в задачах с полными данными. Существенно, что для всех рассмотренных наборов данных предпочтительный подход к сведению многоклассовой задачи к бинарным можно определить при редуцированной размерности пространства признаков. Таким образом, выбор оптимального подхода и функции потерь для конкретной задачи может проводиться при сильном сжатии пространства признаков, что значительно снижает вычислительные затраты.

Литература

[1] *Fisher R. A.* The use of multiple measurements in taxonomic problems // *Annals of Eugenics*, 1936. Vol. 7, No. 7. P. 179–188.

[2] *Cortes C., Vapnik V.* Support-vector networks // *Machine Learning*, 1995. Vol. 20, No. 3. P. 273–297. Available at: `http://dx.doi.org/10.1023/A:1022627411411`.

[3] *Xia F.* Advanced statistical methods in nlp: Multi-class classification. 2012. Available at: `http://courses.washington.edu/ling572/winter2012/slides/ling572_class13_multiclass.pdf`.

[4] *Rifkin R.* Lecture on multiclass classification. 2008. Available at: `http://www.mit.edu/~9.520/spring08/Classes/multiclass.pdf`.

[5] *Tax D. M. J., Duin R. P. W.* Using two-class classifiers for multiclass classification // ICPR. Vol. 2. 2002. P. 124–127. Available at: `http://dx.doi.org/10.1109/ICPR.2002.1048253`.

[6] *Dietterich T. G., Bakiri G.* Solving multiclass learning problems via error-correcting output codes // *Journal of Artificial Intelligence Research*, 1995. Vol. 2. P. 263–286.

[7] *Allwein E. L., Schapire R. E., Singer Y.* Reducing multiclass to binary: A unifying approach for margin classifiers // *Journal of Machine Learning Research*, 2000. Vol. 1. P. 113–141.

[8] *Escalera S., Pujol O., Radeva P.* Error-correcting ouput codes library // *Journal of Machine Learning Research*, 2010. Vol. 11. P. 661–664. Available at: `http://doi.acm.org/10.1145/1756006.1756026`.

[9] *Freund Y., Schapire R.* A short introduction to boosting // *Journal of Japanese Society for Artificial Intelligence*, 1999. Vol. 14, No. 5. P. 771–780. Available at: `http://citeseer.nj.nec.com/freund99short.html`.

[10] *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting // *Journal of Computer and System Sciences*, 1997. Vol. 55, No. 1. P. 119–139. Available at: http://www.sciencedirect.com/science/article/pii/S002200009791504X.

[11] *Pearson K.* On lines and planes of closest fit to systems of points in space // *Philosophical Magazine*, 1901. Vol. 2. P. 559–572.

[12] *Golub G. H., Van Loan C. F.* Matrix Computations. 2nd edition. Baltimore: Johns Hopkins University Press, 1989.

[13] *Goel N., Bebis G., Nefian A.* Face recognition experiments with random projection // Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference / Ed. by A. K. Jain, N. K. Ratha. Vol. 5779. 2005. P. 426–437.

[14] *Dasgupta S., Gupta A.* An elementary proof of a theorem of johnson and lindenstrauss // *Random Struct. Algorithms*, 2003. Vol. 22, No. 1. P. 60–65. Available at: http://dx.doi.org/10.1002/rsa.10073.

[15] *Li P., Hastie T. J., Church K. W.* Very sparse random projections // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06. New York, NY, USA: ACM, 2006. P. 287–296. Available at: http://doi.acm.org/10.1145/1150402.1150436.

[16] *Achlioptas D.* Database-friendly random projections: Johnson-lindenstrauss with binary coins // *J. Comput. Syst. Sci.*, 2003. Vol. 66, No. 4. P. 671–687.

[17] *LeCun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // *Proceedings of the IEEE*, 1998. Vol. 86, No. 11. P. 2278–2324. MNIST database available at http://yann.lecun.com/exdb/mnist/.

[18] *Chang C.-C., Lin C.-J.* LIBSVM: A library for support vector machines // *ACM Transactions on Intelligent Systems and Technology*, 2011. Vol. 2, No. 1. P. 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Медиальная ширина фигуры – дескриптор формы изображений*

Л. М. Местецкий
mestlm@mail.ru
Факультет вычислительной математики и кибернетики МГУ, Москва, Ленинские горы, МГУ, 2-й учебный корпус

Задача генерации классификационных признаков для объектов переменной формы, таких например, как фигура человека или животного, состоит в построении дескрипторов формы, которые сохраняют инвариантность при деформации объектов. В статье предлагается концепция построения такого интегрального дескриптора формы фигуры, называемого функцией медиальной ширины. Функция медиальной ширины определяется на основе скелета и радиальной функции фигуры. Скелет фигуры — это множество точек-центров вписанных в фигуру окружностей. Радиальная функция фигуры определена в точке скелета и равна радиусу вписанной в фигуру окружности с центром в этой точке. По определению медиальная ширина фигуры в точках скелета равна радиальной функции. Предлагается понятие медиальной ширины фигуры в каждой её точке. Ширину фигуры в точке определяем как длину проходящего через эту точку радиуса одного из максимальных вписанных в фигуру кругов. Затем определяем в фигуре подмножество заданной ширины, состоящее из всех точек фигуры, в которых медиальная ширина не превосходит заданного значения. После этого определяем функцию медиальной ширины фигуры, описывающую площадь подмножества заданной ширины как функцию от параметра ширины. Таким образом, функция медиальной ширины представляет собой функцию распределения медиальной ширины в точках фигуры. В статье предлагается эффективный алгоритм вычисления функции медиальной ширины для многоугольной фигуры. Алгоритм основан на построении диаграммы Вороного линейных сегментов, образующих границу фигуры. Алгоритм обобщается для так называемой циркулярной фигуры, получаемой скруглением углов в многоугольной фигуре. Выбор класса циркулярных фигур обусловлен тем, что ими можно аппроксимировать сложные формы объектов растровых изображений. Работоспособность и эффективность предлагаемого подхода демонстрируется вычислительным экспериментом на примере задачи сравнения формы ладоней при биометрической идентификации личности.

Ключевые слова: медиальная ширина; скелет; бицикл; диаграмма Вороного; многоугольная фигура.

Medial width of a figure – an image shape descriptor*

L. M. Mestetskiy
MSU Faculty of Computational Mathematics and Cybernetics, Moscow, Leninskie Gory, MSU, 2nd Education Building

The problem of features generation for classification of flexible objects of variable shape, for example a human figure or an animal figure, is to build shape descriptors which remain invariant during deformation of objects. The paper proposes the concept of building such an integral figure shape descriptor called the function of the medial width. The concept of the

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-01-00716.

medial width function is defined on the basis of the skeleton and the radial function of a figure. The skeleton of a figure is a set of centers of circles inscribed in the figure. The radial function of a figure is defined at the skeleton point and is equal to the radius of the inscribed circle centered at that point.

The medial width in skeleton points is, by definition, equal to the radial function. A concept of the medial width in each point of a figure is introduced. The medial width in a point of a figure is defined as the maximum length of the radius of an inscribed circle passing through the point. The figure's subset of a given width is then defined, consisting of all the points of the figure with medial width not exceeding the given value. After that the medial width function of a figure describing the area of the subset of a given width as a function of the width parameter is defined. Thus, the medial width function is a width distribution function of a figure.

The paper proposes an efficient algorithm to compute the medial width function for polygonal figures. The algorithm is based on the construction of the Voronoi diagram of line segments forming the boundary of the figure. The solution is generalized for the so-called circular figure obtained by replacing the corners of a polygonal figure with conjugate circular arcs. The choice of the class of circular figures is governed by their ability to approximate complex shapes of image objects. Efficiency of the proposed approach is demonstrated by the example of palm shape comparison for biometric identification.

Keywords: medial width; skeleton; bicircle; Voronoi diagram; polygonal figure.

Введение

Классификация формы объектов изображения требует построения признакового описания объектов, отражающего особенности их формы. Задача генерации классификационных признаков для объектов переменной формы, таких например, как фигура человека или животного, состоит в построении дескрипторов формы, которые сохраняют инвариантность при деформации объектов.

Важную роль в классификации формы объектов, имеющих протяжённые элементы, играют срединные оси, или скелеты. Скелет фигуры - это множество точек-центров вписанных в фигуру окружностей. Скелет имеет вид плоского геометрического графа. Анализ этого графа даёт возможность строить различные топологические и метрические признаки формы объекта.

Также информативным признаком формы является ширина объекта относительно срединных осей. Ширина объекта описывается радиальной функцией скелета, которая каждой точке скелета ставит в соответствие радиус вписанной в фигуру окружности с центром в этой точке. Совокупность срединных осей и радиальной функции называется медиальным представлением фигуры [1].

Распределение ширины объекта часто является инвариантом при деформациях объектов, а также имеет отличительные особенности для различных классов объектов. Поэтому ширина объектов используется для генерации признаков при распознавании формы изображений. В качестве интегрального дескриптора ширины объекта может быть использовано понятие pattern spectrum, введённое в [2]. В русскоязычной литературе обычно используется введенный Ю.В. Визильтером термин «морфологический спектр». Аппарат морфологических спектров активно развивается и используется в приложениях к распознаванию изображений в работах Ю.В. Визильтера и учеников [3, 4, 5]. Вообще морфологические спектры имеют различные применения в анализе и распознавании изображений.

Для оценки ширины объекта используется простой частный случай спектра с дисковым структурным элементом.

Традиционный подход к вычислению морфологических спектров основан на использовании методов дискретной математической морфологии. Для получения спектра требуется многократное преобразование изображения с помощью операций морфологического открытия. Поскольку открытия должны выполняться строго последовательно, соответствующие вычисления являются весьма ресурсоемкими по времени. Высокие затраты времени на вычисление морфологического спектра долгое время служили препятствием для его использования при обработке видеопоследовательностей и при анализе сложных изображений высокого разрешения. Работы Ю.В. Визильтера и С.В. Сидякина [4, 5] позволили существенно сократить затраты времени на вычисление морфологических спектров с дисковым структурным элементом за счёт использования концепции непрерывного скелета бинарного изображения [6]. Полученное на основе скелета непрерывное медиальное представление формы объекта позволило сократить большое число операций над растровым изображением, что привело к существенному ускорению вычислений. Однако совсем исключить растровую обработку авторам не удалось. Место растровых операций открытия заняли операции дискретизации скелета и кругов медиального представления объекта. Для этого потребовалась растеризация скелета с помощью алгоритма Брезенхэма и многократная растеризации перекрывающихся кругов. Работа Е.Ю. Макаровой [7] показала, что использование непрерывного медиального представления формы не позволяет точно вычислить морфологический спектр, а допускает лишь некоторое приближенное его построение.

Предлагаемый в настоящей статье новый дескриптор ширины объекта основан на непрерывном медиальном представлении фигуры и может быть вычислен точно с высокой эффективностью. Новый дескриптор, названный функцией медиальной ширины фигуры, основывается на следующих принципах.

1. На основе медиального представления фигуры вводим понятие медиальной ширины фигуры в каждой её точке.
2. Определяем в фигуре подмножество заданной ширины, состоящее из всех точек фигуры, в которых медиальная ширина не превосходит заданное значение.
3. Определяем функцию медиальной ширины фигуры, описывающую площадь подмножества заданной ширины как функцию от параметра ширины.

Функция медиальной ширины является монотонной кусочно-непрерывной. Она может иметь лишь конечное число разрывов первого рода. Эта функция может быть использована для генерации признакового описания формы аналогично морфологическому спектру. Для этого строится ее дискретная разностная гистограмма, которую по аналогии будем называть медиальным спектром фигуры.

В статье предлагается метод прямого вычисления медиальной ширины для многоугольных фигур (многоугольник с многоугольными дырами). Метод основан на построении диаграммы Вороного линейных сегментов, составляющих границу многоугольной фигуры. Многоугольная фигура выбрана потому что, с одной стороны, для построения ее диаграммы Вороного существуют высокоэффективные алгоритмы [8, 6, 9]. С другой стороны, фигуры с нелинейной границей, а также растровые дискретные изображения можно с высокой точностью аппроксимировать многоугольными фигурами. Для более адекватной аппроксимации нелинейных и дискретных фигур вводится понятие циркулярной фигуры. Циркулярная фигура получается в результате процесса стрижки скелета

(pruning), приводящего к «скруглению» углов многоугольной фигуры дугами окружностей. Предложенный метод прямого вычисления медиальной ширины для многоугольных фигур легко обобщается на циркулярные фигуры.

Реализация и экспериментальная оценка предложенного подхода выполнена применительно к задаче биометрической идентификации личности по форме ладони. Известна работа, в которой решение этой задачи выполнено с использованием морфологического спектра [10]. Опыт этот можно считать положительным с точки зрения применения для распознавания формы интегрального показателя ширины объекта. Однако низкая вычислительная эффективность полученного решения не позволяет рассчитывать на его использование в биометрических системах реального времени. Наше решение наряду с хорошим качеством распознавания обеспечивает высокую вычислительную эффективность.

Медиальное представление и медиальная ширина фигуры

Определение 1. *Фигурой называется ограниченная замкнутая область на евклидовой плоскости.*

Определение 2. *Пустым кругом фигуры называется круг, целиком лежащий в фигуре.*

Определение 3. *Пустой круг называется вписанным кругом фигуры, если он является максимальным, т.е. не содержится ни в каком другом пустом круге.*

Определение 4. *Скелетом фигуры называется множество точек-центров всех вписанных кругов фигуры.*

На рис. 1 представлен пример фигуры в виде замкнутой области с одной дырой, граница которой состоит из двух замкнутых контуров. Здесь же показаны примеры вписанных кругов и скелет фигуры. Следует отметить особо круги A и B , имеющие нулевой радиус.

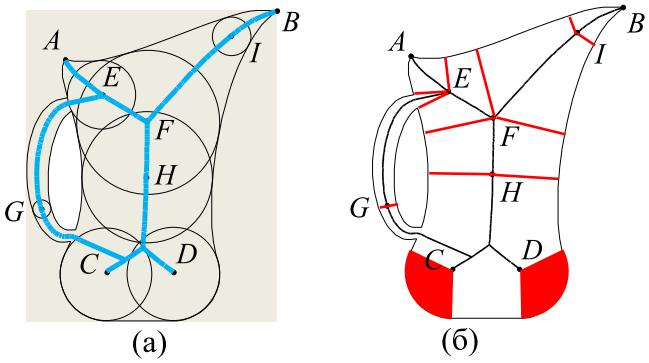


Рис. 1. (а) фигура, вписанные круги, скелет, (б) спицы.

Определение 5. *Радиальная функция определена в точках скелета и ставит в соответствие точке скелета радиус вписанного круга с центром в этой точке.*

Определение 6. *Спицей называется отрезок прямой, соединяющий точку скелета с ближайшей точкой границы фигуры.*

Количество спиц, связанных с точкой скелета, может существенно различаться. В примере на рис. 1б показаны спицы для некоторых точек скелета. Точки A и B имеют по одной спице нулевой длины, точки G, H, I имеют по две спицы, точки E и F – по три. А для точек C и D имеется бесконечное число спиц, которые заполняют сектора.

Будем использовать следующие обозначения:
R^2 – евклидова плоскость,
G – фигура, т.е. ограниченная замкнутая область $G \subset R^2$,
∂G – граница фигуры G,
G' – внутренняя открытая область фигуры $G' = G \backslash \partial G$,
$C(P)$ – пустой круг с центром в точке $P \in G$,
S – скелет фигуры G.

Отметим некоторые важные свойства, связывающие точки фигуры с множеством спиц этой фигуры.

Лемма 1. *Через каждую точку фигуры проходит хотя бы одна спица. Следовательно, спицы покрывают всю фигуру.*

Доказательство. Пусть $P \in G$ и $Q \in \partial G$ ближайшая к P точка на границе фигуры (рис. 2). Пустой круг $C(P)$ с центром P имеет радиус PQ. Если этот круг максимальный, то P точка скелета,$P \in S$, и тогда отрезок PQ является спицей, инцидентной точке P. Если круг $C(P)$ не максимальный, то существует круг $C(P_1)$, такой что $C(P) \in C(P_1)$. Тогда $Q \in C(P_1)$ и круги $C(P)$ и $C(P_1)$ имеют общую касательную в точке Q, а точка P_1 лежит на прямой PQ. В силу ограниченности фигуры G среди таких пустых кругов $C(P_1)$ существует максимальный круг $C(P^*)$. Тогда отрезок P^*Q является спицей, и точка P лежит на этой спице.

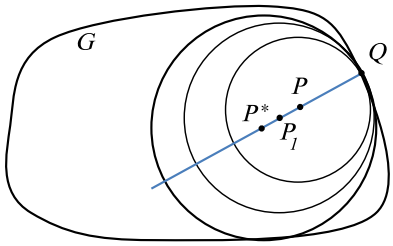


Рис. 2. К лемме 1.

Заметим, что данное рассуждение подходит и для частного случая, когда точка P является граничной, $P \in \partial G$, и $P = Q$. ■

Лемма 2. *Через каждую внутреннюю точку фигуры, не являющуюся точкой скелета, проходит только одна спица.*

Доказательство. Рассмотрим внутреннюю точку $P \in G'$ (рис. 3). Предположим, что через неё проходит несколько спиц. Выберем две из них AB и CD. Пусть A и C точки скелета, $A, C \in S$, а B и D – точки границы, $B, D \in \partial G$. Из $P \notin S$ и $P \notin \partial G$ следует, что P есть внутренняя точка отрезков AB и CD.

Треугольники $\triangle APD$ и $\triangle CPB$ невырожденные, следовательно $AP + PD > AD$ и $CP + PB > BC$.

Рассмотрим отрезки PB и PD.

Если $PB \geqslant PD$ то $AB = AP + PB \geqslant AP + PD > AD$. А если $PB < PD$ то $CD = = CP + PD > CP + PB > CB$. Но точки A и C являются центрами максимальных пустых кругов, поэтому $AB \leqslant AD$ и $CD \leqslant CB$.

Следовательно, существование нескольких спиц, проходящих через точку P, невозможно. ■

Теперь определим понятие медиальной ширины фигуры в точке.

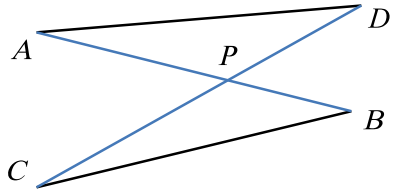


Рис. 3. К лемме 2.

Определение 7. *Медиальная ширина фигуры во внутренней точке равна длине минимальной спицы, инцидентной этой точке.*

Из леммы 1 следует, что медиальная ширина определена для всех внутренних точек фигуры. Все спицы, инцидентные одной точке скелета, имеют одинаковую длину равную радиусу пустого круга с центром в этой точке. Поэтому для точек скелета медиальная ширина просто равна радиальной функции. Для каждой внутренней точки, не являющейся точкой скелета, существует согласно лемме 2 единственная инцидентная спица, поэтому медиальная ширина в такой точке определена однозначно. Граничные точки фигуры могут иметь несколько инцидентных спиц.

Обозначим

$\varphi(g), g \in G'$ – медиальная ширина фигуры в точке g,

$G'_w = \{g \in G', \varphi(g) \leqslant w\}$ – подмножество точек фигуры, в которых медиальная ширина не превосходит заданное значение $w \geqslant 0$,

$\mathcal{F}(w) = \mu(G'_w)$ – функция медиальной ширины фигуры, площадь множества точек G'_w,

$f(w) = \frac{d\mathcal{F}(w)}{dw}$ – медиальный спектр фигуры.

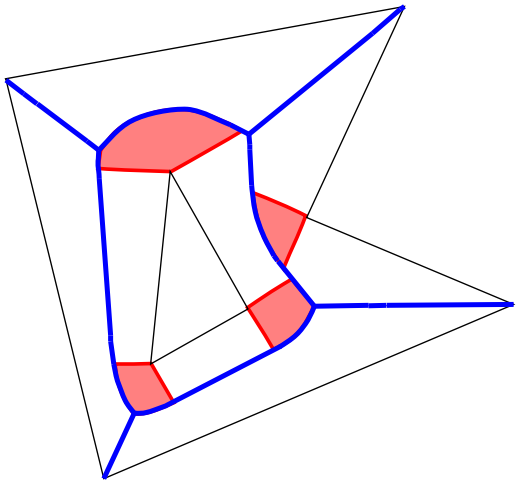


Рис. 4. Точки на границе многоугольной фигуры, имеющие более одной инцидентной спицы.

Точки фигуры, имеющие несколько инцидентных спиц разной длины, могут находиться лишь в угловых точках в вогнутых вершинах на границе фигуры. Примеры таких точек приведены на рис. 4. Но общая площадь границы равна нулю. Поэтому точки границы не вносят вклад в вычисление площади области заданной ширины. Поэтому медиальную ширину в точках границы можно положить равной нулю.

На рис. 5 представлен пример, показывающий подмножество точек фигуры, в которых медиальная ширина не превосходит заданное значение.

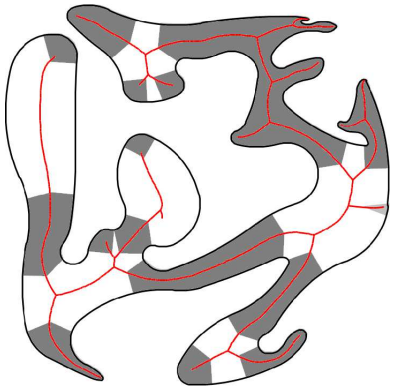


Рис. 5. Область заданной ширины в фигуре.

Многоугольные и циркулярные фигуры и их медиальная ширина

Многоугольная фигура определяется как ограниченная замкнутая область с границей, состоящей из многоугольников. Многоугольными фигурами можно аппроксимировать с высокой точностью любые фигуры, граница которых задана жордановыми кривыми. Кроме того, их можно использовать в качестве удобной непрерывной модели для аппроксимации объектов бинарных растровых изображений.

Границу многоугольной фигуры можно представить в виде множества сайтов-точек (вершин фигуры) и сайтов-сегментов (сторон фигуры). Для этого множества сайтов определена так называемая диаграмма Вороного линейных сегментов. Мы будем называть часть этой диаграммы Вороного, лежащую внутри фигуры, диаграммой Вороного (ДВ) многоугольной фигуры.

ДВ многоугольной фигуры представляет собой геометрический граф, рёбрами которого являются отрезки прямых линий и квадратичных парабол.

Пусть G – многоугольная фигура, $Vor(G) = \langle V, E\rangle$ – ДВ фигуры G. Здесь V – множество вершин, E – множество рёбер ДВ. С каждым ребром ДВ связана пара сайтов, для которых линия ребра является бисектором – общей границей их ячеек Вороного. Рассмотрим подграф ДВ $\langle V, E\rangle$, образованный из $Vor(G)$ путём отсечения части терминальных вершин и рёбер, инцидентных этим вершинам. Как известно, если отсечь вершины и рёбра $Vor(G)$, инцидентные вогнутым вершинам многоугольной фигуры, то объединение рёбер полученного подграфа образует скелет фигуры, т.е. $S = \bigcup_{e\in E'} e$. Это позволяет рассматривать скелет многоугольной фигуры как подграф ДВ $S = \langle V', E'\rangle$, $V' \subseteq V$, $E' \subseteq E$.

Пусть S скелет многоугольной фигуры G. Определим процесс дальнейшей стрижки скелета, как последовательное отсечение некоторых терминальных вершин и инцидентных им рёбер ДВ, т.е. построение подграфов $S_1, S_2, \ldots, S_n$ таких, что $S_i = \langle V_i, E_i\rangle$, $S_{i+1} = \langle V_{i+1}, E_{i+1}\rangle$, $V_{i+1} = V_i \setminus \{v_i\}$, $E_{i+1} = E_i \setminus \{e_i\}$, $v_i \in V_i$, $e_i \in E_i$, и при этом вершина v_i является терминальной в подграфе S_i и она имеет инцидентное ребро e_i.

Определение 8. *Подграфы ДВ, получаемые в результате процесса стрижки, называются скелетными подграфами.*

Определение 9. *Объединение $G' = \bigcup_{P\in S'} C(P)$ пустых кругов, центры которых расположены на скелетном подграфе $S' \subseteq S$, называется силуэтом скелетного подграфа или циркулярной фигурой.*

Многоугольная фигура G может быть представлена в виде объединения всех пустых кругов с центрами в точках скелета $G = \bigcup_{P \in S} C(P)$, т.е. она является силуэтом полного скелета и представляет собой частный случай циркулярной фигуры.

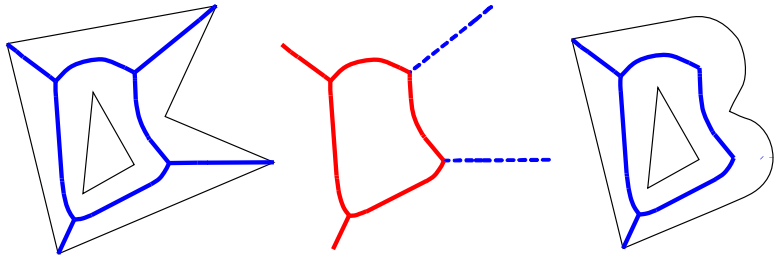


Рис. 6. (а) многоугольная фигура и ее скелет, (б) скелетный подграф, полученный в результате стрижки, (в) циркулярная фигура.

В примере на рис. 6 представлена многоугольная фигура, её скелет, скелетный подграф, полученный в результате стрижки, и циркулярная фигура, образованная оставшимися после стрижки пустыми кругами этого подграфа.

Бициклы в многоугольных и циркулярных фигурах

Определение 10. *Бициклом ребра скелета $e \in E$ называется объединение всех вписанных кругов с центрами на ребре e.*

Определение 11. *Собственной областью называется объединение всех спиц, инцидентных точкам ребра e и линейным ребрам границы фигуры.*

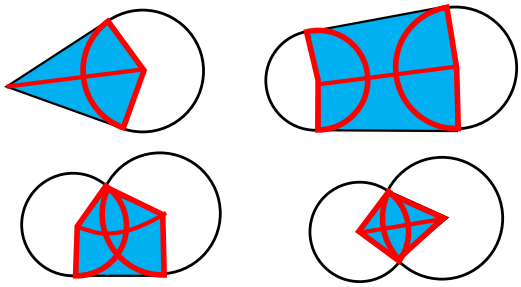


Рис. 7. Бициклы, собственные области, внешние и внутренние сектора концевых кругов.

Собственная область лежит целиком в бицикле. Граница собственной области состоит из двух сайтов и двух спиц. Круги с центрами в вершинах, инцидентных ребру, называются концевыми кругами бицикла. Каждый концевой круг разбивается на два сектора - внутренний и внешний. Внешний сектор опирается на дугу концевого круга, входящую в границу бицикла, а внутренний сектор представляет собой дополнение внешнего сектора в концевом круге (рис. 7).

Бициклы образуют покрытие всей многоугольной фигуры. При этом собственные области бициклов могут пересекаться только по своим граничным спицам.

Пусть B^e – собственная область бицикла ребра e. Рассмотрим множество точек $g \in B^e_x$, в которых медиальная ширина не превосходит величину $x \geqslant 0$, т.е. $B^e_x =$

$= \{g \in B^e, \varphi(g) \leqslant x\}$. Это множество B^e_x будем называть областью ширины x. Обозначим $\mu\,(B^e_x)$ – площадь этой области.

Определение 12. *Функцией медиальной ширины бицикла называется зависимость* $\mathcal{F}^e(x) = \mu\,(B^e_x)$ *площади области ширины* x *от параметра* x.

Собственные области двух рёбер могут иметь непустое пересечение по граничным спицам. Площадь пересечения равна нулю. Очевидно, что собственные области всех рёбер скелета покрывают всю многоугольную фигуру, поэтому функция медиальной ширины фигуры может быть выражена в виде суммы функций медиальной ширины бициклов рёбер её скелетного графа $\mathcal{F}(x) = \sum_{e\in E} \mathcal{F}^e(x)$.

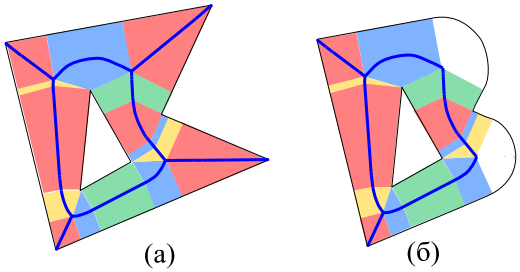


Рис. 8. Покрытие многоугольной и циркулярной фигур собственными областями бициклов и граничными секторами.

Циркулярная фигура также представляет собой объединение бициклов всех рёбер своего скелетного графа. Но при этом собственные области этих бициклов не покрывают целиком циркулярную фигуру. Часть циркулярной фигуры оказывается покрытой только внешними секторами бициклов (рис. 8б). С каждой вершиной скелета $v \in V$ связан один или несколько бициклов инцидентных рёбер.

Определение 13. *Граничным сектором циркулярной фигуры будем называть часть вписанного круга с центром в вершине* $v \in V$, *которая не покрывается собственными областями бициклов фигуры.*

Обозначим $\theta(v)$ площадь граничного сектора вершины v. Для циркулярной фигуры функция медиальной ширины выражается через медиальную ширину бициклов и площади граничных секторов

$$\mathcal{F}(x) = \sum_{e\in E} \mathcal{F}^e(x) + \sum_{v\in V, r_v \leqslant x(v)} \theta(v).$$

Это позволяет свести задачу вычисления функции медиальной ширины многоугольных и циркулярных фигур к вычислению медиальной ширины отдельных бициклов.

Вычисление медиальной ширины бициклов

Различаются три типа бициклов в зависимости от пары образующих сайтов: сегмент-сегмент, сегмент-точка и точка-точка. В бицикле сегмент-сегмент ось бицикла является прямой линией, такой бицикл называется линейным. В бицикле сегмент-точка ось бицикла является параболой, поэтому такой бицикл будем называть параболическим.

Введём для бицикла с сайтами точка-точка местную систему прямоугольных декартовых координат, в которой начало координат расположено посредине между сайтами-точками, а ось ординат проходит через сайты-точки. Центры пустых кругов этого би-

цикла лежат на оси абсцисс. Пусть q – расстояние между сайтами. Тогда радиус пустого круга с центром в точке $(x, 0)$ определяется как $\rho = \sqrt{\left(\frac{q}{2}\right)^2 + x^2}$. Это уравнение в системе координат (x, ρ) описывают гиперболу, поэтому бицикл такого типа будем называть гиперболическим.

Мы хотим получить в явном виде формулы для вычисления медиальной ширины $\mathcal{F}_{lin}(z)$, $\mathcal{F}_{par}(z)$, $\mathcal{F}_{hyp}(z)$ для всех трёх типов бициклов в виде зависимости от параметра ширины z. При этом бицикл задаётся только лишь радиусами своих концевых кругов r и R и расстоянием между центрами этих кругов l.

Медиальная ширина линейного бицикла. На рис. 9 представлен линейный бицикл и связанные с ним точки.

AB – ось бицикла, A, B – центры концевых кругов,

A_1B_1, A_2B_2 – проекции оси бицикла на сайты-сегменты.

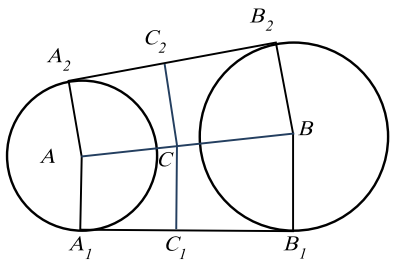


Рис. 9. Линейный бицикл.

Лемма 3. *Медиальная ширина линейного бицикла описывается уравнением*

$$\mathcal{F}_{lin}(z) = \begin{cases} 0 & \text{при } z < r \\ az^2 + b & \text{при } r \leqslant z \leqslant R \\ t(R + r) & \text{при } z > R \end{cases} \tag{1}$$

где

$$a = \begin{cases} 0 & \text{при } r = R \\ \frac{t}{R-r} & \text{при } r < R \end{cases} \tag{2}$$

$$b = \begin{cases} 2lr & \text{при } r = R \\ -\frac{tr^2}{R-r} & \text{при } r < R \end{cases} \tag{3}$$

$$t = \sqrt{l^2 - (R - r)^2}.$$

Доказательство. Рассмотрим случай $r < R$. Если $z \in [r, R]$, то на оси бицикла найдётся точка C, являющаяся центром пустого круга с радиусом z. Пусть C_1 и C_2 проекции C на сайты-сегменты. Тогда область ширины z есть многоугольник $A_1AA_2C_2CC_1$ и значение $\mathcal{F}_{lin}(z)$ равно площади $\mu\,(A_1AA_2C_2CC_1)$. А эта площадь складывается из суммы площадей одинаковых трапеций $\mu\,(A_1ACC_1)$ и $\mu\,(AA_2C_2C)$.

Площадь трапеции

$$\mu\,(A_1ACC_1) = \frac{(AA_1 + CC_1) \cdot A_1C_1}{2}.$$

Из пропорции $\frac{A_1C_1}{CC_1-AA_1} = \frac{A_1B_1}{BB_1-AA_1}$ получаем

$$A_1C_1 = \frac{A_1B_1}{BB_1 - AA_1} \cdot (CC_1 - AA_1)\,.$$

Поскольку $AA_1 = r$, $CC_1 = x$, $A_1B = t$, имеем

$$\mu\,(A_1ACC_1) = \frac{r+z}{2} \cdot \frac{t \cdot (z-r)}{(R-r)} = \frac{t}{2 \cdot (R-r)} \cdot \left(z^2 - r^2\right).$$

Отсюда

$$\mathcal{F}_{lin}(x) = 2 \cdot \mu\,(A_1ACC_1) = \frac{t}{(R-r)} \cdot z^2 - \frac{t \cdot r^2}{(R-r)},$$

что даёт значения коэффициентов в (2) и (3) при условии $r < R$.

При $z < r$ очевидно, что $\mathcal{F}_{lin}(z) = 0$, а в случае $z > R$ получаем

$$\mathcal{F}_{lin}(z) = \mathcal{F}_{lin}(R) = \frac{t}{(R-r)} \cdot (R^2 - r^2) = t \cdot (R+r).$$

Теперь рассмотрим случай $r = R$, когда концевые круги бицикла имеют одинаковый радиус. В этом случае функция медиальной ширины имеет скачок в точке z. Величина скачка определяется площадью многоугольника $A_1AA_2C_2CC_1$, который в данном случае вырождается в прямоугольник $A_1A_2C_2C_1$ со сторонами $A_1C_1 = l$ и $A_1A_2 = 2r = R + r$. Поэтому

$$\mathcal{F}_{lin}(z) = \begin{cases} 0 & \text{при } z < r \\ 2tr & \text{при } z \geqslant r \end{cases}$$

что доказывает формулы (1)–(3) при $r = R$. ■

Медиальная ширина параболического бицикла. На рис. 10 представлен параболический бицикл.

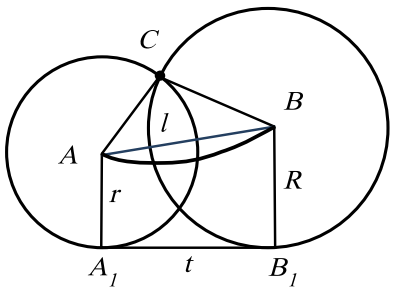


Рис. 10. Параболический бицикл.

AB – отрезок параболы – ось бицикла, A, B – центры концевых кругов,
A_1, B_1 – проекции оси бицикла на сайт-сегмент,
C – сайт-точка параболического бицикла,
r, R – радиусы концевых кругов бицикла, $r \leqslant R$,
l – расстояние между их центрами,
$t = \sqrt{l^2 - (R-r)^2}$ – длина проекции оси бицикла на директрису параболы.

Ось бицикла представляет собой отрезок параболы. При этом вершина параболы может лежать как на оси бицикла или вне её. Если вершина параболы лежит на оси бицикла, особо выделим случай, когда она совпадает с концевой точкой оси. В этом случае центр меньшего концевого круга совпадает с вершиной параболы и находится в макушке параболы. Такой бицикл будем называть корневым. Положение вершины параболы относительно оси параболического бицикла определяется следующим условием.

Лемма 4. *Если* $t \geqslant 2 \cdot \sqrt{r \cdot (R-r)}$, *то вершина параболы лежит на оси бицикла, если* $t < 2 \cdot \sqrt{r \cdot (R-r)}$, *то вне оси.*

Доказательство. Рассмотрим частный случай, когда бицикл является корневым и макушка параболы совпадает с центром его концевого круга (рис. 11).

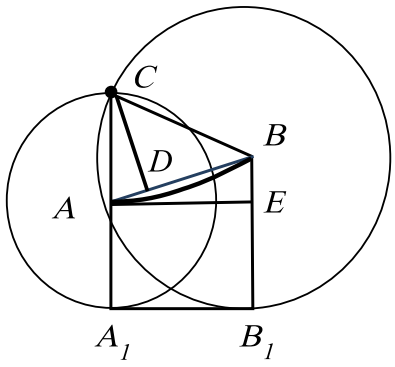


Рис. 11. Корневой параболический бицикл (к леммам 4 и 6).

Точка D – проекция точки C на AB, точка E – проекция точки A на BB_1. В треугольнике $\triangle ABC$ $AB = l, AC = r, BC = R$. В треугольнике $\triangle ACD$ $DC^2 = AC^2 - AD^2$, в треугольнике $\triangle BCD$ $DC^2 = BC^2 - BD^2$. Следовательно,

$$AC^2 - AD^2 = BC^2 - BD^2. \tag{4}$$

Пусть длина $AD = y$. Тогда из (4) следует

$$r^2 - y^2 = R^2 - (l-y)^2 \text{ и}$$

$$2ly = l^2 - R^2 + r^2. \tag{5}$$

Треугольники $\triangle ABE$ и $\triangle ACD$ подобны, поэтому $\frac{AC}{AD} = \frac{AB}{BE}$. Отсюда $\frac{r}{y} = \frac{l}{R-r}$ и $y = \frac{r(R-r)}{l}$. Подставляя это значение y в (5), получаем

$$2 \cdot r \cdot (R-r) = l^2 - R^2 + r^2$$

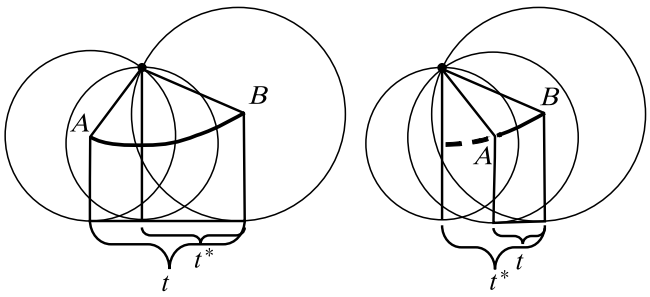
$$2rR - 2r^2 - l^2 + R^2 - r^2 = 0 \tag{6}$$

Подставляем также $l^2 = t^2 + (R-r)^2$ в (6):

$$2rR - 3r^2 - t^2 - (R-r)^2 + R^2 = 0$$

$$4rR - 4r^2 - t^2 = 0$$

$$t^2 = 4r(R-r)$$



(а) (б)

Рис. 12. (а) вершина параболы лежит на оси бицикла, (б) вершина вне оси бицикла.

$$t^* = 2\sqrt{r(R-r)}. \tag{7}$$

Формула (7) описывает длину проекции оси бицикла для рассматриваемого частного случая, когда макушка параболы совпадает с центром концевого круга бицикла. Рассмотрим теперь общий случай, когда такого совпадения нет.

Если вершина параболы лежит на оси бицикла, то очевидно, имеет место соотношение $t > t^*$, а если вне оси, то $t < t^*$. Это следует из следующих геометрических соображений. В случае, когда вершина лежит на оси бицикла (рис. 12а), можно сдвинуть меньший круг A по направлению к большему и найти такое положение меньшего круга, при котором бицикл превращается в корневой (рис. 11). В этом случае $t > t^*$. Если же вершина параболы лежит вне оси (рис. 12б), то сдвиг меньшего круга A осуществляем в направлении от большего круга B, и тогда $t < t^*$. Это и доказывает утверждение леммы. Заметим, что в случае, когда размеры концевых кругов одинаковы, вершина параболы всегда лежит внутри оси бицикла. ■

Выберем систему декартовых координат для параболического ребра, чтобы вершина параболы имела координаты $(0,0)$, фокус параболы (сайт-точка) с координатами $C\left(0, \frac{p}{2}\right)$, а директриса параболы – прямая $y = -\frac{p}{2}$ (ей принадлежит сайт-сегмент).

В этой системе координат уравнение параболы имеет вид $y = \frac{1}{2p}x^2$.

Параметр параболы p для параболического бицикла определяется на основе следующей леммы.

Лемма 5. *Параметр параболы оси параболического бицикла есть*

$$p = \frac{t^2}{2l^2}\left(R + r + \sqrt{(R+r)^2 - l^2}\right).$$

Доказательство. Пусть $(x_1, y_1), (x_2, y_2)$ координаты центров меньшего и большего концевых кругов бицикла. Поскольку центры концевых кругов лежат на параболе, координаты связаны уравнением параболы, т.е. $y_1 = \frac{1}{2p}x_1^2$ и $y_2 = \frac{1}{2p}x_2^2$. К тому же эти точки равноудалены от сайта-точки C и сайта-сегмента A_1B_1, поэтому $y_1 = r - \frac{p}{2}$ и $y_2 = R - \frac{p}{2}$. Отсюда имеем

$$\frac{1}{2p}x_1^2 = r - \frac{p}{2}, \tag{8}$$

$$\frac{1}{2p}x_1^2 = R - \frac{p}{2}. \tag{9}$$

Вычитая (8) из (9), получаем $x_2^2 - x_1^2 = 2p(R-r)$ и, поскольку $x_2 = t + x_1$, то

$$t \cdot (2x_1 + t) = 2p \cdot (R - r).$$

Отсюда $x_1 = \frac{1}{2}\left(\frac{2p\cdot(R-r)}{t} - t\right)$.

Подставляем это выражение в уравнение (8) и получаем уравнение для определения p:

$$\left(\frac{(R-r)^2}{t^2} + 1\right)\cdot p^2 - (R+r)\cdot p + \frac{1}{4}t^2 = 0.$$

Это квадратное уравнение имеет два корня

$$p = \frac{t^2}{2}\cdot\frac{(R+r) \pm \sqrt{4Rr - t^2}}{(R-r)^2 + t^2}.$$

Подставляем $t^2 = l^2 - (R-r)^2$, тогда эта формула упрощается

$$p = \frac{t^2}{2\cdot l^2}\left(R + r \pm \sqrt{(R+r)^2 - l^2}\right). \tag{10}$$

Два корня этого уравнения соответствуют двум точкам пересечения концевых окружностей бицикла. Одна из этих точек совпадает с сайтом-точкой бицикла, а другая лежит внутри бицикла. Очевидно, что сайт-точка соответствует большему значения параметра из (10). Поэтому из двух корней (10) следует выбрать большее значение

$$p = \frac{t^2}{2\cdot l^2}\left(R + r + \sqrt{(R+r)^2 - l^2}\right).$$

■

Теперь на основе полученного выражения для параметра параболического бицикла появляется возможность вычислить его медиальную ширину.

Рассмотрим корневой бицикл, у которого ось есть сегмент параболы с параметром p. Тогда радиус меньшего концевого круга равен $\frac{p}{2}$. Пусть радиус большего концевого круга равен z. Нас интересует площадь собственной области этого бицикла.

Лемма 6. *Площадь собственной области параболического корневого бицикла с параметром p и радиусом концевого круга z равна*

$$\Phi(z) = (z+p)\sqrt{\frac{p}{2}\left(z - \frac{p}{2}\right)}.$$

Доказательство. Пусть больший круг корневого бицикла имеет центр в точке $B(x, y)$ и радиус z (рис. 11). Собственная область бицикла имеет форму трапеции A_1CBB_1. Площадь этой трапеции $\Phi(z) = \frac{1}{2}(A_1C + BB_1)\cdot A_1B_1$.

Поскольку $A_1C = p, BB_1 = z, A_1B_1 = x$, получаем

$$\Phi(z) = \frac{1}{2}(p+z)\cdot x. \tag{11}$$

Из уравнения параболы $y = \frac{1}{2p}x^2$ и условия $CB^2 = x^2 + \left(y - \frac{p}{2}\right)^2 = z^2$ имеем

$$x^2 + \left(\frac{1}{2p}x^2 - \frac{p}{2}\right)^2 = z^2$$

Преобразуя левую часть этого уравнения, получаем $\left(\frac{1}{2p}x^2 + \frac{p}{2}\right)^2 = z^2$.

Отсюда $\frac{1}{2p}x^2 + \frac{p}{2} = z$ и

$$x = \sqrt{2pz - p^2} = \sqrt{p(2z - p)}.$$

Подставляя в (11), получаем искомую формулу

$$\Phi(z) = (z + p)\sqrt{\frac{p}{2}\left(z - \frac{p}{2}\right)}.$$

■

Следствие 1. *Функция медиальной ширины корневого параболического бицикла с параметром p и концевым кругом радиуса R есть*

$$\Phi(z, p, R) \begin{cases} 0 & \text{если } z \leqslant \frac{p}{2} \\ \varphi(z) & \text{если } \frac{p}{2} < z \leqslant R. \\ \varphi(R) & \text{если } z > R \end{cases}$$

Функцию медиальной ширины любого (не корневого) параболического бицикла можно теперь вычислить на основе площадей корневых бициклов.

Лемма 7. *Медиальная ширина параболического бицикла с радиусами концевых кругов r, R и параметром параболы p имеет следующий вид:*

При положении вершины параболы на оси

$$\mathcal{F}_{par}(z) = \Phi(z, p, r) + \Phi(z, p, R)$$

При положении вершины параболы вне оси

$$\mathcal{F}_{par}(z) = \Phi(z, p, R) - \Phi(z, p, r)$$

Доказательство. Бицикл с концевыми кругами (r, R) можно представить как композицию двух корневых бициклов с кругами $\left(\frac{p}{2}, r\right)$ и $\left(\frac{p}{2}, R\right)$ (рис. 13). Если вершина параболы лежит на оси бицикла, то собственная область бицикла есть объединение собственных областей этих корневых бициклов. А если вершина параболы лежит вне оси, то собственная область есть замыкание разности собственных областей корневых бициклов (рис. 13). В соответствии с этим, функция медиальной ширины этого бицикла складывается как сумма или разность функций медиальной ширины корневых бициклов. ■

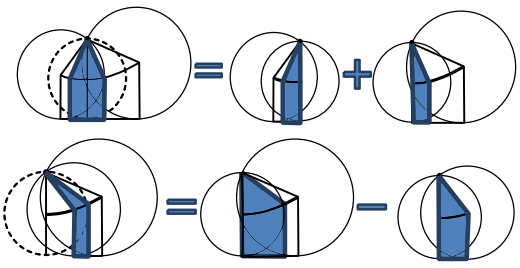


Рис. 13. Медиальная ширина параболического бицикла (Лемма 7).

Таким образом, леммы 4-7 дают формулы для вычисления медиальной ширины параболического бицикла по параметрам r, R, l. Сначала нужно определить положение вершины параболы относительно оси бицикла (лемма 4), затем найти параметр параболы оси

(лемма 5), и после этого можно вычислить значение функции и медиальной ширины в любой точке (лемма 6, 7).

Медиальная ширина гиперболического бицикла. На рис. 14 представлен гиперболический бицикл. Здесь отрезок AB – это ось бицикла, r, R $(r \leqslant R)$ – радиусы концевых кругов с центрами в точках A и B, l – расстояние между центрами концевых кругов, C, D – образующие сайты-точки бицикла. Точку пересечения прямой CD с прямой AB будем называть центром гиперболического бицикла.

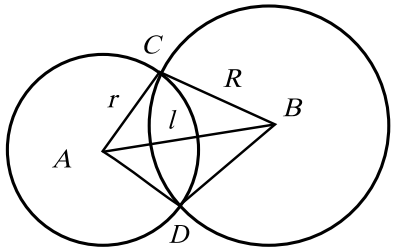


Рис. 14. Гиперболический бицикл.

В зависимости от соотношения величин r, R, l точки A и B могут находиться по одну сторону от прямой CD или по разные стороны от неё. В первом случае центр лежит вне оси бицикла, а во втором - на оси. Положение центра в гиперболическом бицикле определяется следующим условием.

Лемма 8. *Если $l^2 + r^2 \geqslant R^2$, то центр лежит на оси бицикла, а если $l^2 + r^2 < R^2$, то вне оси.*

Доказательство. Пусть E – центр бицикла (точка пересечения прямых AB и CD). Если E совпадает с центром концевого круга A, т.е. $A = E$, то $BE^2 + CE^2 = BC^2$ или $l^2 + r^2 = R^2$.

Если центр лежит внутри оси бицикла (рис. 15), т.е. точка E находится внутри отрезка AB, то $AC^2 + AB^2 > CE^2 + BE^2 = BC^2$. Поскольку при этом $AC = r, AB = l, BC = R$, получаем $l^2 + r^2 > R^2$. Если же центр E лежит вне оси бицикла, то это значит что точка A лежит внутри отрезка BE (на рисунке это положение обозначено точкой A'). Тогда имеет место $A'C^2 + A'B^2 < CE^2 + BE^2 = BC^2$. Поскольку в этом случае $A'C = r, A'B = l, BC = R$, получаем $l^2 + r^2 < R^2$. ■

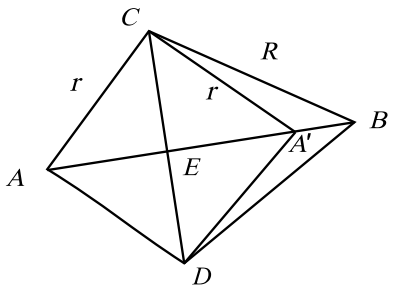


Рис. 15. Определение положения вершины гиперболического бицикла.

Будем называть параметром гиперболического бицикла расстояние q между его сайтами-точками.

Лемма 9. *Параметр гиперболического бицикла с радиусами концевых кругов r, R и длиной оси l равен $q = \frac{1}{l}\sqrt{[(l+R)^2 - R^2] \cdot [R^2 - (l-r)^2]}$.*

Доказательство. Параметр q равен длине отрезка CD на рис. 15. Тогда

$$AE = \sqrt{AC^2 - CE^2} = \sqrt{r^2 - \frac{q^2}{4}}$$

$$BE = \sqrt{BC^2 - CE^2} = \sqrt{R^2 - \frac{q^2}{4}}$$

В случае, когда проекция сайтов-точек лежит на оси бицикла, имеем $AE + BE = l$, т.е.

$$\sqrt{r^2 - \frac{q^2}{4}} + \sqrt{R^2 - \frac{q^2}{4}} = l.$$

Простыми преобразованиями получаем

$$l - \sqrt{r^2 - \frac{q^2}{4}} = \sqrt{R^2 - \frac{q^2}{4}}$$

$$l^2 + \left(r^2 - \frac{q^2}{4}\right) - 2l \cdot \sqrt{r^2 - \frac{q^2}{4}} = R^2 - \frac{q^2}{4}$$

$$\frac{l^2 + r^2 - R^2}{2l} = \sqrt{r^2 - \frac{q^2}{4}}$$

$$\left(\frac{l^2 + r^2 - R^2}{2l}\right)^2 = r^2 - \frac{q^2}{4} \tag{12}$$

$$q^2 = 4\left[r^2 - \left(\frac{l^2 + r^2 - R^2}{2l}\right)^2\right] = \frac{1}{l^2}[4l^2r^2 - (l^2 + r^2 - R^2)] =$$

$$= \frac{1}{l^2}(2lr + l^2 + r^2 - R^2)(2lr - l^2 - r^2 + R^2) = \frac{1}{l^2}[(l + R)^2 - R^2][-(l - R)^2 + R^2].$$

Отсюда и следует искомая формула

$$q = \frac{1}{l}\sqrt{[(l + R)^2 - R^2] \cdot [R^2 - (l - r)^2]}. \tag{13}$$

В случае, когда проекция сайтов лежит вне оси бицикла, получаем $BE - A'E = l$, т.е.

$$\sqrt{R^2 - \frac{q^2}{4}} - \sqrt{r^2 - \frac{q^2}{4}} = l.$$

В этом случае выполняем аналогичные преобразования

$$l + \sqrt{r^2 - \frac{q^2}{4}} = \sqrt{R^2 - \frac{q^2}{4}}$$

$$l^2 + r^2 - \frac{q^2}{4} + 2l\sqrt{r^2 - \frac{q^2}{4}} = R^2 - \frac{q^2}{4}$$

$$2l\sqrt{r^2 - \frac{q^2}{4}} = R^2 - r^2 - l^2$$

$$r^2 - \frac{q^2}{4} = \left(\frac{R^2 - r^2 - l^2}{2l}\right). \tag{14}$$

Уравнение (14) преобразуется точно так же, как и уравнение (12). Из него получаем такое же выражение для параметра q, что и в (13).

Таким образом, формула из утверждения леммы справедлива для обоих случаев. ■

По аналогии с разобранным ранее случаем параболического бицикла назовём гиперболический бицикл корневым, если центр его концевого круга совпадает с вершиной бицикла.

Лемма 10. *Площадь собственной области гиперболического корневого бицикла с параметром q и радиусом концевого круга z равна*

$$\Psi(z) = \frac{q}{2}\sqrt{z^2 - \frac{q^2}{4}}.$$

Доказательство. Собственная область корневого гиперболического бицикла представляет собой треугольник $\triangle BCD$, площадь которого равна $S = \frac{1}{2} \cdot CD \cdot AB$ (рис. 16). Из $CD = q$ и $AB = \sqrt{BC^2 - AC^2} = \sqrt{z^2 - \left(\frac{q}{2}\right)^2}$ получаем

$$\Psi(z) = S = \frac{1}{2} \cdot CD \cdot AB = \frac{1}{2} \cdot q \cdot \sqrt{z^2 - \left(\frac{q}{2}\right)^2}.$$

■

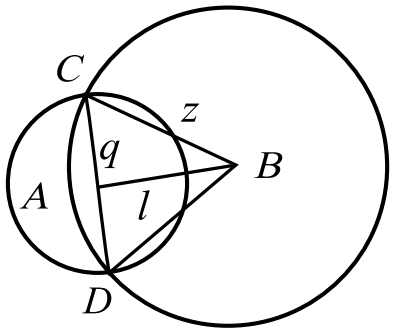


Рис. 16. Определение площади собственной области корневого гиперболического бицикла (к лемме 10).

Следствие 2. *Функция медиальной ширины корневого гиперболического бицикла с параметром q и концевым кругом радиуса R есть*

$$\Psi(z, p, R) \begin{cases} 0 & \text{если } z \leqslant \frac{p}{2} \\ \psi(z) & \text{если } \frac{p}{2} < z \leqslant R. \\ \psi(R) & \text{если } z > R \end{cases}$$

Вычисление медиальной ширины гиперболического бицикла осуществляется через площади собственных областей двух корневых гиперболических бициклов, построенных на его основе.

Лемма 11. *Медиальная ширина гиперболического бицикла с радиусами концевых кругов r, R и параметром q имеет следующий вид:*

При положении центра на оси бицикла

$$\mathcal{F}_{hyp}(z) = \Psi(z, q, r) + \Psi(z, q, R)$$

При положении центра вне оси бицикла

$$\mathcal{F}_{hyp}(z) = \Psi(z, q, R) - \Psi(z, q, r)$$

Доказательство. Доказывается аналогично лемме 7. Рассматриваем два корневых гиперболических бицикла, у которых собственные области образуют собственную область исходного бицикла (рис. 17). Корневые бициклы имеют общий концевой круг, центром которого является центр бицикла, а радиус равен $\frac{q}{2}$. Если центр принадлежит оси исходного бицикла, то площадь собственной области бицикла равна сумме площадей собственных областей его корневых бициклов. А если центр лежит вне оси, то площадь собственной области бицикла равна разности площадей собственных областей корневых бициклов. В соответствии с этим, функция медиальной ширины этого бицикла складывается как сумма или разность функций корневых бициклов. ■

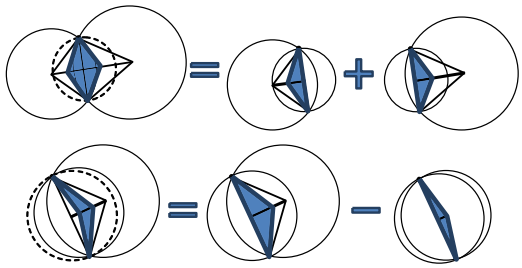


Рис. 17. Медиальная ширина гиперболического бицикла (Лемма 11).

Медиальная ширина концевых секторов бицикла. Для вычисления функции медиальной ширины циркулярной фигуры необходимо определить площади секторов концевых кругов бициклов, которые не покрываются собственными областями этих бициклов.

Концевой круг бицикла разбивается на две части, представляющие собой сектора этого круга. Эти сектора соответствуют двум дугам окружности концевого круга - внутренней и внешней. Внутренняя дуга покрывается другими кругами бицикла, а внешняя дуга образует границу бицикла.

С каждой вершиной скелета $v \in V$ связано несколько бициклов, соответствующих инцидентным рёбрам этой вершины $e_1, e_2, \ldots, e_k, k \geqslant 1$. Концевые круги этих бициклов с центром в точке v совпадают. Внутренние сектора этих бициклов не перекрываются, поэтому если угловые размеры внутренних секторов бициклов равны $\alpha_1, \alpha_2, \ldots, \alpha_k$, то $\alpha_1 + \alpha_2 + \cdots + \alpha_k \leqslant 2\pi$. Если в вершине v сохранились все рёбра, которые входят в скелет многоугольной фигуры, то $\alpha_1 + \alpha_2 + \cdots + \alpha_k = 2\pi$. А в тех вершинах циркулярной фигуры, в которых в процессе стрижки были отсечены какие-то рёбра скелета, сумма внутренних углов меньше, т.е. $\alpha_1 + \alpha_2 + \cdots + \alpha_k < 2\pi$. Получается, что угловой размер секторов вершины v, не покрытых внутренними секторами бициклов, равен

$$\xi(v) = 2\pi - (\alpha_1 + \alpha_2 + \cdots + \alpha_k). \tag{15}$$

Если r_v - радиус пустого круга вершины v, то площадь внешних секторов вершины v есть

$$\theta(v) = \frac{1}{2}\xi(v) \cdot (r_v)^2.$$

Для вычисления $\xi(v)$ для всех вершин скелета $v \in V$, нужно вычислить размеры внутренних секторов концевых кругов бициклов, образующих циркулярную фигуру.

Угловые размеры внутренних секторов бициклов определяются в зависимости от их типа (линейный, параболический или гиперболический), размеров концевых кругов $r, R, (r \leqslant R)$ и расстояния l между центрами этих кругов.

Лемма 12. *Внутренняя дуга концевого круга линейного бицикла имеет размер* $\alpha = \pi + 2 \cdot \arcsin \frac{R-r}{l}$ *для меньшего круга и* $\alpha = \pi - 2 \cdot \arcsin \frac{R-r}{l}$ *для большего круга.*

Доказательство. Внутренняя дуга (рис. 18) меньшего концевого круга A измеряется углом

$$\alpha = \angle A_1AA_2 = \angle A_1AC_1 + \angle C_1AC_2 + \angle C_2AA_2 =$$
$$= \frac{\pi}{2} + 2 \cdot \angle C_1AB + \frac{\pi}{2} = \pi + 2 \cdot \arcsin \frac{BC_2}{AB}.$$

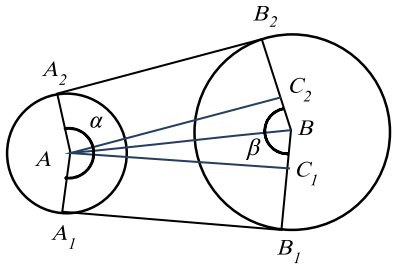


Рис. 18. Внутренняя дуга линейного бицикла (Лемма 12).

Поскольку $BC_2 = R - r$, а $AB = l$, получается формула для меньшего круга $\alpha = \pi + 2 \cdot \arcsin \frac{R-r}{l}$.

Размер внутренней дуги большего концевого круга получается из соотношения

$$\beta = 2\pi - \alpha = \pi - 2 \cdot \arcsin \frac{R-r}{l}.$$

■

Лемма 13. *Внутренняя дуга большего концевого круга параболического бицикла с параметром* p *равна* $\alpha = \arccos\left(1 - \frac{p}{R}\right)$, *а внутренняя дуга меньшего круга равна* $\beta = \arccos\left(1 - \frac{p}{r}\right)$ *в случае, когда вершина параболы лежит на оси бицикла, и* $\beta = 2\pi - \arccos\left(1 - \frac{p}{r}\right)$, *если вне оси.*

Доказательство. В полярной системе координат (ρ, φ) с центром в вершине параболы и осью, совпадающей с осью ординат, уравнение параболы имеет вид $\rho = \frac{p}{1-\cos\varphi}$. В точке параболы, являющейся центром большего круга $\rho = R$, а в центре меньшего круга $\rho = r$. Внутренние дуги концевых кругов бицикла имеют размеры в интервале от 0 до 2π. Уравнение $\rho = \frac{p}{1-\cos\varphi}$ даёт в этом интервале два решения $\varphi_1 = \arccos\left(1 - \frac{p}{\rho}\right)$ и $\varphi_2 = 2\pi - \arccos\left(1 - \frac{p}{\rho}\right)$. Одно из этих значений меньше, а другое больше π.

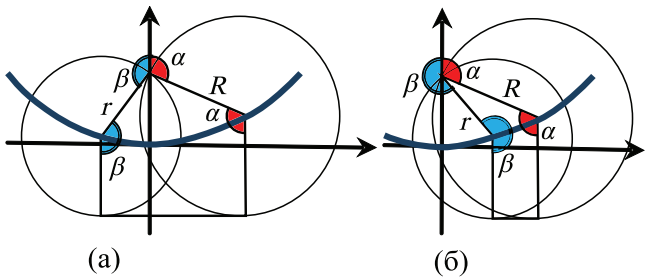


Рис. 19. Внутренняя дуга параболического бицикла (Лемма 13).

Из геометрических соображений ясно, что внутренняя дуга большего круга всегда меньше π, поэтому её величина есть $\alpha = \arccos\left(1 - \frac{p}{R}\right)$. А для меньшего круга возможны два варианта решения. Если вершина параболы лежит внутри оси бицикла (рис. 19а), то внутренняя дуга меньшего концевого круга тоже меньше π, и поэтому $\beta = \arccos\left(1 - \frac{p}{r}\right)$. А в случае, когда вершина параболы лежит вне оси бицикла (рис. 19б), величина этой дуги больше π. Тогда $\beta = 2\pi - \arccos\left(1 - \frac{p}{r}\right)$. ■

Лемма 14. *Внутренняя дуга большего концевого круга гиперболического бицикла с параметром q имеет размер $\alpha = \arcsin\left(\frac{q}{2R}\right)$, а внутренняя дуга меньшего круга имеет размер $\beta = \arcsin\left(\frac{q}{2r}\right)$ в случае, когда вершина бицикла лежит на оси бицикла и $\beta = 2\pi - \arcsin\left(\frac{q}{2r}\right)$, если вне оси.*

Доказательство. Очевидно следует из приведенного рис. 20. ■

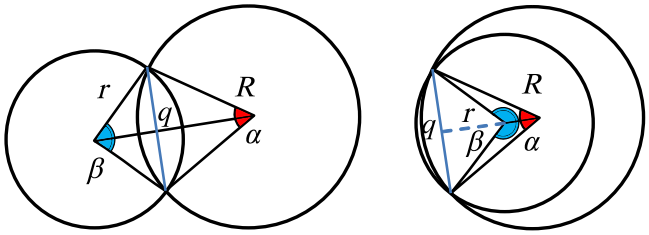


Рис. 20. Внутренние дуги гиперболического бицикла (Лемма 14).

Алгоритм вычисления медиальной ширины фигуры

Входом алгоритма служит медиальное представление многоугольной или циркулярной фигуры, имеющее формат взвешенного графа.

$S = \langle V, E \rangle$ – скелет фигуры,
V – множество вершин графа,
E – множество рёбер графа,
$e = (r_e, R_e, l_e, k_e)$ – ребро графа,
r_e – радиус меньшего круга ребра,
R_e – радиус большего круга ребра,
l_e – расстояние между центрами кругов ребра,
k_e – тип ребра (lin — линейное, par – параболическое, hyp – гиперболическое).
$v = (r_v, c_v)$ – вершина графа,
r_v – радиус круга вершины,
c_v – открытая дуга вершины (вычисляется в ходе работы алгоритма).
Результат работы алгоритма:

$\{F_0, F_1, \ldots, F_N\}$ – массив значений функции распределения медиальной ширины фигуры $F_i = \mathcal{F}(r_i)$ в некотором заранее определённом наборе значений аргумента $\{r_0, r_1, \ldots, r_N\}$. В качестве такого набора значений аргумента может быть использована, например, последовательность целых чисел $\{0, 1, \ldots, N\}$, т.е. $r_i = i$.

// Функция вычисления площади собственной области параболического корневого бицикла с параметром p и радиусом концевого круга z

Функция $\Phi(z)$

$\Phi(z) = (z+p)\sqrt{\frac{p}{2}\left(z - \frac{p}{2}\right)}$

// Функция вычисления площади собственной области гиперболического корневого бицикла с параметром q и радиусом концевого круга z

Функция $\Psi(z)$

$\Psi(z) = \frac{q}{2}\sqrt{z^2 - \frac{q^2}{4}}$

// Функция вычисления медиальной ширины линейного бицикла

Функция $\mathcal{F}_{lin}(x)$

$$\mathcal{F}_{lin}(x) = \begin{cases} 0 & \text{при } x < r \\ ax^2 + b & \text{при } r \leqslant x \leqslant R \\ t(R+r) & \text{при } x > R \end{cases}$$

// Функция вычисления медиальной ширины параболического бицикла

Функция $\mathcal{F}_{par}(z)$

если $TopPos$ **то**

$$\mathcal{F}_{par}(z) = \begin{cases} 0 & \text{при } z \leqslant \frac{p}{2} \\ 2 \cdot \Phi(z) & \text{при } \frac{p}{2} < z \leqslant r \\ \Phi(r) + \Phi(z) & \text{при } r < z \leqslant R \\ \Phi(r) + \Phi(R) & \text{при } z > R \end{cases}$$

иначе

$$\mathcal{F}_{par}(z) = \begin{cases} 0 & \text{при } z \leqslant \frac{p}{2} \\ \Phi(z) - \Phi(r) & \text{при } r < z \leqslant R \\ \Phi(R) - \Phi(r) & \text{при } z > R \end{cases}$$

// Функция вычисления медиальной ширины гиперболического бицикла

Функция $\mathcal{F}_{hyp}(z)$

если $TopPos$ **то**

$$\mathcal{F}_{hyp}(z) = \begin{cases} 0 & \text{при } z \leqslant \frac{q}{2} \\ 2 \cdot \Psi(z) & \text{при } \frac{q}{2} < z \leqslant r \\ \Psi(r) + \Psi(z) & \text{при } r < z \leqslant R \\ \Psi(r) + \Psi(R) & \text{при } z > R \end{cases}$$

иначе

$$\mathcal{F}_{hyp}(z) = \begin{cases} 0 & \text{при } z \leqslant r \\ \Psi(z) - \Psi(r) & \text{при } r < z \leqslant R \\ \Psi(R) - \Psi(r) & \text{при } z > R \end{cases}$$

// Начало программы

для $i \in \{0, 1, \ldots, N\}$

$F_i \leftarrow 0$

для всех $e \in E$

// предобработка, подготовка параметров для всех бициклов
если $k_e = lin$ **то**
$t = \sqrt{l_e^2 - (R_e - r_e)^2}$
если $r_e = R_e$ **то**
$a \leftarrow 0$
$b \leftarrow 2 l_e r_e$
иначе
$a \leftarrow \frac{t}{R_e - r_e}$
$b \leftarrow -\frac{t r_e^2}{R_e - r_e}$
$i_{min} \leftarrow \max\{i | r_i \leqslant r_e\}$
$i_{max} \leftarrow \min\{i | r_i \geqslant R_e\}$
иначе если $k_e = par$ **то**
$t = \sqrt{l_e^2 - (R_e - r_e)^2}$
$TopPos \leftarrow (t \geqslant 2\sqrt{r \cdot (R - r)})$ *// положение вершины параболы относительно оси параболич. бицикла: TRUE – на оси, FALSE – вне оси*
$p = \frac{t^2}{2l^2}(R + r + \sqrt{(R + r)^2 - l^2})$ *// параметр параболического бицикла*
если $TopPos$ **то**
$i_{min} \leftarrow \max\left\{i | r_i \leqslant \frac{p}{2}\right\}$
иначе
$i_{min} \leftarrow \max\{i | r_i \leqslant r_e\}$
$i_{max} \leftarrow \min\{i | r_i \geqslant R_e\}$
иначе если $k_e = hyp$ **то**
$TopPos \leftarrow (l^2 + r^2 \geqslant R^2)$ *// положение центра гиперболического бицикла: TRUE – на оси, FALSE – вне оси*
$q = \frac{1}{l}\sqrt{[(l + r)^2 - R^2] \cdot [R^2 - (l - r)^2]}$ *// параметр гиперболического бицикла*
если $TopPos$ **то**
$i_{min} \leftarrow \max\left\{i | r_i \leqslant \frac{q}{2}\right\}$
иначе
$i_{min} \leftarrow \max\{i | r_i \leqslant r_e\}$
$i_{max} \leftarrow \min\{i | r_i \geqslant R_e\}$
// подготовка параметров для концевых секторов
$v_1 \leftarrow e.v_1$ *// Вершина - конец ребра e, соответствующий меньшему кругу бицикла*
$v_2 \leftarrow e.v_2$ *// Вершина - конец ребра e, соответствующий большему кругу бицикла*
если $k_e = lin$ **то**
$v_1.c \leftarrow v_1.c + \pi + 2 \cdot \arcsin\frac{R_e - r_e}{l_e}$
$v_2.c \leftarrow v_2.c + \pi - 2 \cdot \arcsin\frac{R_e - r_e}{l_e}$
иначе если $k_e = par$ **то**
$v_2.c \leftarrow v_2.c + \arccos\left(1 - \frac{p}{R_e}\right)$
если $TopPos$ **то**
$v_1.c \leftarrow v_1.c + \arccos\left(1 - \frac{p}{r_e}\right)$
иначе
$v_1.c \leftarrow v_1.c + 2\pi - \arccos\left(1 - \frac{p}{r_e}\right)$
иначе если $k_e = hyp$ **то**

$v_2.c \leftarrow v_2.c + \arcsin\left(\frac{q}{2R_e}\right)$
если $TopPos$ **то**
$v_1.c \leftarrow v_1.c + \arcsin\left(\frac{q}{2r_e}\right)$
иначе
$v_1.c \leftarrow v_1.c + 2\pi - \arcsin\left(\frac{q}{2r_e}\right)$
// вычисление вклада бицикла в общую функцию медиальной ширины фигуры
для $i = i_{min} + 1$ до i_{max}
$r \leftarrow r_i$
если $k_e = lin$ **то**
$H_i \leftarrow H_i + \mathcal{F}_{lin}(r_i) - \mathcal{F}_{lin}(r_{i-1})$
иначе если $k_e = par$ **то**
$H_i \leftarrow H_i + \mathcal{F}_{par}(r_i) - \mathcal{F}_{par}(r_{i-1})$
иначе если $k_e = hyp$ **то**
$H_i \leftarrow H_i + \mathcal{F}_{hyp}(r_i) - \mathcal{F}_{hyp}(r_{i-1})$
// вычисление вклада концевых секторов в общую функцию медиальной ширины фигуры
для всех $v \in V$
$j \leftarrow \min\{i | r_i \geqslant vc_v.r\}$
$H_j \leftarrow H_j + (2\pi - v.c) \cdot \frac{(v.r)^2}{2}$
// вычисление функции медиальной ширины в заданных значениях аргумента
$F_0 \leftarrow 0$
для $i = 1$ до N
$F_i \leftarrow F_{i-1} + H_i$
// конец алгоритма

Реализация и вычислительный эксперимент

Алгоритм реализован в среде Дельфи-Паскаль на базе алгоритма построения непрерывного скелета бинарного изображения. Общая структура алгоритма включает в себя следующие этапы:

— построение аппроксимирующей многоугольной фигуры для бинарного растрового изображения. Аппроксимирующая фигура имеет границу, состоящую из одного или нескольких многоугольников. Многоугольники являются разделяющими многоугольниками минимального периметра. При этом многоугольники задают подразбиение плоскости на связные компоненты, обладающие следующим свойством. Все пиксели растрового изображения, являющиеся внутренними точками одной и той же компоненты, имеют одинаковый цвет. Пиксели, лежащие на границе компоненты, могут иметь разные цвета. Алгоритм построения аппроксимирующей многоугольной фигуры описан в [6];
— построение внутренней части диаграммы Вороного для множества линейных сегментов, составляющих границу аппроксимирующей многоугольной фигуры. Построение осуществляется с помощью алгоритма, описанного в [6];
— построение аппроксимирующей циркулярной фигуры на основе выделения скелетного графа из диаграммы Вороного и его последующей регуляризации;
— вычисление функции медиальной ширины для полученной циркулярной фигуры с помощью алгоритма, описанного выше.

Пример на рис. 21 иллюстрирует описанную схему.

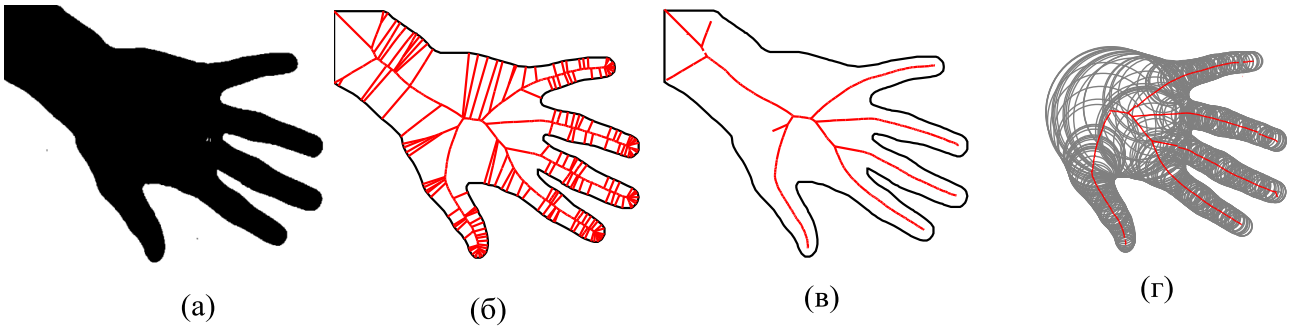


Рис. 21. Аппроксимация изображения циркулярной фигурой.

Исходное бинарное изображение представлено битовой картой 640x480 пикселей (рис 21а). Аппрксимирующая многоугольная фигура имеет вид простого многоугольника с 346 вершинами. Его скелетный граф имеет 689 рёбер (рис. 21б). Простая регуляризация с параметром стрижки 1 оставляет в скелетном графе 435 рёбер (рис. 21в). Полученный в результате стрижки подграф обладает следующим свойством. Объединение всех вписаных в исходную фигуру кругов с центрами на этом подграфе образуют циркулярную фигуру, которая отличается от многоугольной фигуры не более чем на заданную величину ε в метрике Хаусдорфа. В нашем примере $\varepsilon = 1$. Далее семантическая сегментация оставляет в скелетном графе только ту его часть, которая описывает пальцы и пясть. В результате получается граф с 382 рёбрами. Этот граф задаёт циркулярную фигуру, состоящую из 382 бициклов (рис. 21г). Среди них 182 линейных, 152 параболических и 48 гиперболических. Общий вид функции медиальной ширины представлен на диаграмме рис. 22.

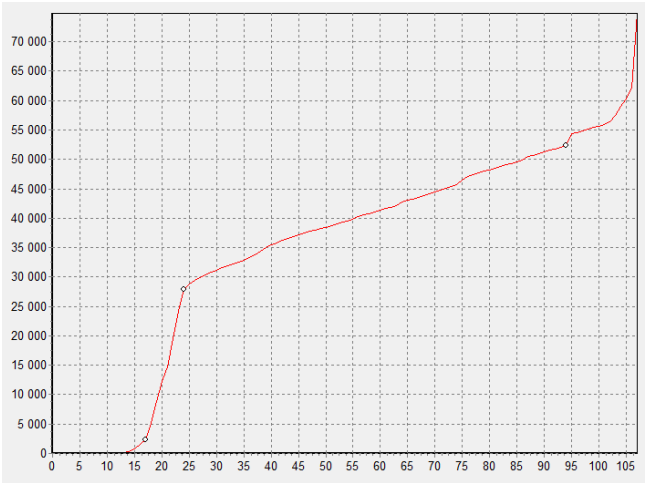


Рис. 22. График функции медиальной ширины ладони.

Оценка скорости работы предложенного алгоритма получена на задаче биометрической идентификации личности по геометрии ладони. В качестве меры близости изображений ладоней I_1 и I_2, имеющих функции медиальной ширины $\mathcal{F}_1(x)$ и $\mathcal{F}_2(x)$, принято расстояние между функциями медиальными ширины изображений в метрике L_1

$$\sigma(I_1, I_2) = \int_0^{\infty} \left| \lambda_1^2 \cdot \mathcal{F}_1\left(\frac{1}{\lambda_1} \cdot x\right) - \lambda_2^2 \cdot \mathcal{F}_2\left(\frac{1}{\lambda_2} \cdot x\right) \right| dx. \tag{16}$$

Рис. 23. Фрагмент базы изображений ладоней.

Коэффициенты λ_1 и λ_2 предназначены для нормализации функций медиальной ширины, поскольку изображения ладоней получены при различной дальности съёмки и имеют разный масштаб. Эти коэффициенты вычисляются специальным алгоритмом масштабирования. В эксперименте использованы 160 бинарных изображений ладоней 35 человек (рис. 23). Изображения 640x480 представлены в виде монохромных файлов в формате BMP. Сначала для всех изображений вычисляются медиальные функции ширины. После этого для каждой пары изображений вычисляется расстояние в форме (16). По результатам сравнения строится ROC-кривая, показывающая классификационные возможности данной меры сравнения изображений ладоней (рис. 24).

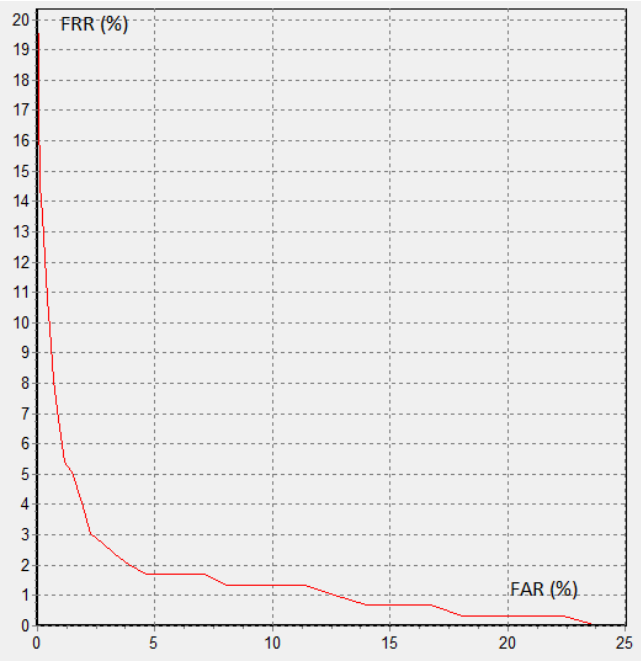


Рис. 24. ROC-кривая сравнения формы ладоней на основании функции медиальной ширины.

В таблице приведены данные о времени вычислений для процессора Intel® Core™ i5-3210M CPU @ 2.50GHz.

Операция	Колич.	Общее время	Время на одну операцию
Вычисление функции медиальной ширины	160 изобр.	2325 мс	14.53 мс
Вычисление парных расстояний	12720 сравн.	3200 мс	0.25 мс

Операция «Вычисление функции медиальной ширины» включает в себя построение аппроксимирующей многоугольной фигуры, вычисление медиального представления, регуляризацию скелета, а также непосредственное вычисление функции медиальной ширины на основе медиального представления.

Заключение

Предложенный метод открывает новые возможности по применению высокоэффективных алгоритмов вычислительной геометрии в анализе и распознавании растровых дискретных изображений. Известные подходы к вычислению дескрипторов для ширины фигур на основе морфологического спектра не подходят для использования в реальном времени работы систем компьютерного зрения, поскольку имеют высокую вычислительную сложность. Предлагаемый переход к непрерывной модели на основе многоугольных фигур, а также высокоэффективный метод вычисления медиальной ширины для таких фигур дают возможность преодолеть этот недостаток, позволяют сравнивать и измерять сходство фигур по их ширине.

Литература

[1] *Siddiqi K., Pizer S. M.* Medial Representations: Mathematics, Algorithms and Applications. Springer, 2008.

[2] *Maragos P.* Pattern Spectrum and Multiscale Shape Representation // *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 1989. Vol. 11, No. 7. P. 701–716.

[3] *Визильтер Ю. В., Сидякин С. В.* Морфологические спектры // *Техническое зрение в системах управления – 2012. Труды научно-технической конференции*, Москва: ИКИ РАН, 2012. С. 234–241.

[4] *Визильтер Ю. В., Сидякин С. В.* Использование морфологических спектров для классификации двумерных фигур и бинарных изображений // *Вестник компьютерных и иформационных технологий.* 2013. № 7. С. 20–28.

[5] *Vizilter Yu., Sidyakin S., Rubis A., Gorbazevich V.* Morphological shape comparison based on skeleton representations // *Pattern Recognition and Image Analysis*, 2012. Vol. 22, No. 3. P. 412–418.

[6] *Местецкий Л. М.* Непрерывная морфология бинарных изображений. Фигуры, скелеты, циркуляры. Москва: Физматлит, 2009.

[7] *Макарова Е. Ю.* Непрерывные алгоритмы морфологического анализа и сравнения листьев растений. Дипломная работа, МГУ, ВМК, 2012.

[8] *Held M.* Vroni and ArcVroni: Software for and Applications of Voronoi Diagrams in Science and Engineering // *Proc. 8th Int. Symp. on Voronoi Diagrams in Science and Engineering (ISVD)*, June 2011. P. 3–12. doi: 10.1109/ISVD.2011.9.

[9] *Karavelas M.* A robust and efficient implementation for the segment Voronoi diagram // *Proc. 1st Int. Symp. on Voronoi Diagrams in Science and Engineering (ISVD)*, Sept. 2004. P. 51–62.

[10] *Ramirez-cortes J. M., Gomez-gil P., Sanchez-perez G., Baez-lopez D.* A Feature Extraction Method Based on the Pattern Spectrum for Hand Shape Biometry // *Proc. World Congress on Engineering and Computer Science*, Oct. 2008.

References

[1] *Siddiqi K., Pizer S. M.* 2008. Medial Representations: Mathematics, Algorithms and Applications. Springer.

[2] *Maragos P.* 1989. Pattern Spectrum and Multiscale Shape Representation. *IEEE Tran. on Pattern Analysis and Machine Intelligence* 11(7):701–716.

[3] *Vizilter Yu. V., Sidyakin S. V.* 2012. Morphological spectra. *Computer vision in control systems 2012. Proceedings of the scientific-technical conference.* Moscow: IKI RAS. 234–241. (in Russ.)

[4] *Vizilter Yu. V., Sidyakin S. V.* 2013. The classification of two-dimensional figures and binary images using morphological pattern spectra. *Herald of computer and information technologies* 7:20–28. (in Russ.)

[5] *Vizilter Yu., Sidyakin S., Rubis A., Gorbazevich V.* 2012. Morphological shape comparison based on skeleton representations. *Pattern Recognition and Image Analysis* 22(3):412–418.

[6] *Mestetskiy L. M.* 2009. Continuous morphology of binary images: Figures, skeletons and circulars. Moscow: Fizmatlit. (in Russ.)

[7] *Makarova E. Yu.* 2012. Continuous algorithms of morphological analysis and comparison of plant leaves. Diploma Thesis, CMC MSU. (in Russ.)

[8] *Held M.* 2011. Vroni and ArcVroni: Software for and Applications of Voronoi Diagrams in Science and Engineering. *Proc. 8th Int. Symp. on Voronoi Diagrams in Science and Engineering (ISVD).* June 2011. 3–12. doi: 10.1109/ISVD.2011.9.

[9] *Karavelas M.* 2004. A robust and efficient implementation for the segment Voronoi diagram. *Proc. 1st Int. Symp. on Voronoi Diagrams in Science and Engineering (ISVD).* Sept. 2004. 51–62.

[10] *Ramirez-cortes J. M., Gomez-gil P., Sanchez-perez G., Baez-lopez D.* 2008. A Feature Extraction Method Based on the Pattern Spectrum for Hand Shape Biometry. *Proc. World Congress on Engineering and Computer Science.* Oct. 2008.

Использование метода ближайших соседей при восстановлении обстановки осадконакопления

Б. В. Белозеров[1]***, А. С. Бочков***[1]***, О. С. Ушмаев***[1]***, О. М. Фукс***[1,2]
Phuks.OM@gazprom-neft.ru
[1]ООО «Газпромнефть НТЦ», Санкт-Петербург, Россия, [2]МФТИ, Москва, Россия

Целью данной работы является построение метода изучения геологической структуры нефтяных месторождений и создание модуля для автоматизации распознавания литолого-фациальной обстановки на основе промысловых данных каротажных диаграмм. Методами машинного обучения решается задача классификации по типам фациальных обстановок каротажных данных с месторождения с использованием спектрального представления геофизических полей. Данная методика была успешно применена к интерпретации фаций на реальном месторождении.

Ключевые слова: геостатистика; обстановка осадконакопления; фациальный анализ; спектральный метод; машинное обучение

Application of nearest neighbour method for sedimentation environment study

B. V. Belozerov[1]***, A. S. Bochkov***[1]***, O. S. Ushmaev***[1]***, and O. M. Fuks***[1,2]
[1]LLC "Gazpromneft NTC", St. Petersburg, Russia, [2]MIPT, Moscow, Russia

Background: Distribution properties of sedimentary rocks determine considerably the geometry and size of the reservoir and, consequently, the volume of hydrocarbon reserves. Therefore knowledge about general patterns of sedimentation rock formation is of crucial practical importance. This work suggests the method to study the geological structure of oil field by automated recognition of lithofacial environment on the basis of geophysical field data.

Methods: In the work, the spectral method is used in geophysical field representation, which is well-known for its effective application to the simulation of low-permeability and high-splitted reservoirs in the nonsteady and anisotropic conditions. The input data are geophysical data interpreted by the filed geologist, which then form the training set for the machine learning algorithm. To reduce the dimensionality of the data, only their significant features (Fourier coefficients) are retained in the learning step of the algorithm. Further, the data are classified into the different facial regions using the machine learning technique.

Results: The method was tested on the real field and with the electrometric well data as the input, it allowed to classify the wells according to the lithofacial sedimentation environment.

Concluding Remarks: In the article, the method of facial environment reconstruction is described and its applicability to the real field is shown.

Keywords: geostatistics; sedimentation environment; facial analysis; spectral method; machine learning

Введение

Как известно, аккумуляция осадков, в которых возможна генерация углеводородов, происходит при определенных физико-географических условиях. Особенности распространения осадочных пород в пространстве и во времени значительно определяют фильтрационно-емкостные свойства залежи (ФЕС), от которых в свою очередь зависят продуктивность и, в конечном счете, накопленная добыча скважин. Поэтому знание общих и частных закономерностей образования осадочных толщ имеет неоспоримое практическое значение. При создании концепции разработки месторождения основным источником информации о залежи являются данные геофизических исследований с разведочных скважин. При прогнозировании распространения свойств геофизических параметров существенным может оказаться знание характера осадконакопления, которое привело к образованию той или иной геологической обстановки. Интерпретация фаций[1] является важным этапом в определении условий седиментации обломочных осадков — сформировалось ли отложение на месте речной дельты, либо в результате миграции русла меандрирующей реки, глубоководное или мелководное это отложение и т. д. Таким образом, комплексная диагностика условий формирования осадков дает ценную информацию, необходимую для построения согласованной геологической модели месторождения и оптимальной эксплуатации залежи.

Построение палеогеографических карт, отражающих наиболее достоверные обстановки осадконакопления на стадии завершения формирования продуктивного пласта-коллектора, представляет собой чрезвычайно науко- и трудоемкую задачу. Для большинства терригенных горных пород широко распространен метод анализа каротажных диаграмм по отдельным скважинам и определения фаций на основе их электрометрических моделей [1]. Но данный вид работ требует не только глубокого понимания основ процесса осадконакопления, но и наличие качественной информации о литологическом составе изучаемых горных пород. К сожалению, наличие качественного описания кернового материала не всегда представляется возможным в виду проведения подобных работ различными компаниями в различное время. К тому же большинство материалов представлены в бумажном виде, что затрудняет и замедляет работу по обработке исходной информации. С целью ускорения и оптимизации работы геолога и предлагается данная методика для распознавания литолого-фациальной обстановки месторождения.

Результативность и прогнозная способность производственных решений при разработке нефтяных месторождений во многом зависит от качества исходной геологической модели. В основе почти всех современных пакетах геологического моделирования лежат классические методы геостатистики [2, 3], главным из которых является вариограммный метод. При этом предположения классической геостатистики — это стационарность и изотропность случайного поля геофизического параметра (а также их ослабления: тренд и стационарные остаток и геометрическая анизотропия) [4]. Однако ни стационарность, ни изотропность в своих строгих формах почти не встречаются при работе с реальными данными. В настоящее время будущее нефтяной отрасли состоит в осваивании низкопроницаемых и высокорасчлененных коллекторов, применение к которым вариограммного метода дает заведомо ложный результат [5]. Поэтому для обеспечения высоких технолого-

[1]Фация — это тело горной породы со специфическими особенностями. При описании осадочной породы фация может быть выделена по цвету, характеру слоистости, составу, структуре, ископаемым остаткам, осадочным текстурам. Отбор признаков для определения фации и вес, присваиваемый каждому из них, зависят от субъективной оценки геолога.

экономических показателей при моделировании таких коллекторов в условиях нестационарности, зональной изменчивости и анизотропности геофизических полей требуются другие методы, лишенные ограничений классической геостатистики. Одним из таких методов является спектральный подход в представлении геофизического поля [5–7]. Как правило, каротажные данные с месторождения представляют собой сильно и случайным образом флуктуирующие функции. Пример каротажных данных приведен на рис. 1.

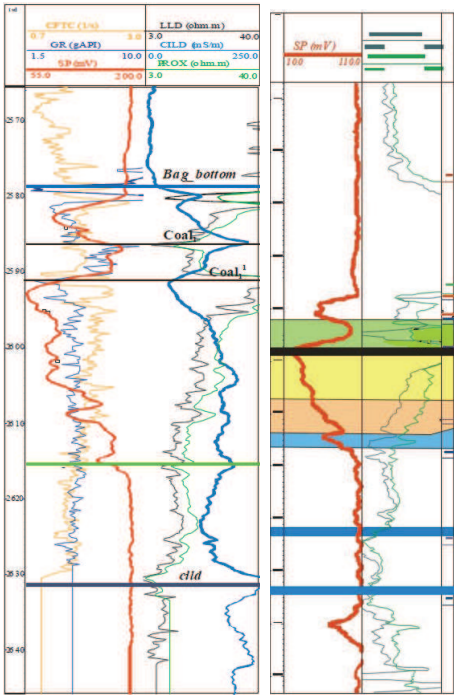


Рис. 1. Пример каротажных данных: по оси y откладывается обычно глубина, по оси x — значения каротажа

Далее статья организована следующим образом. В первом разделе описаны постановка задачи и цель работы. Затем речь идет об используемых методах решения — спектральном методе представлении геофизических полей, выборе информативных признаков для описания каротажных данных и об алгоритме машинного обучения. Последний раздел содержит пример опробования методики для классификации фаций на участке реального месторождения и анализ ошибки предложенного алгоритма.

Постановка задачи

Как было упомянуто ранее, данные геофизических исследований являются основным источником информации о продуктивном пласте на стадии построения геологической модели и создании концепции разработки месторождения. Различные формы каротажных данных свидетельствуют о различных условиях осадконакопления и, соответственно, условиях формирования коллектора. Целью данной работы является разработка метода распознавания литолого-фациальной обстановки по каротажным данным.

В терминах машинного обучения задача распознавания фациальных обстановок относится к задаче классификации — необходимо разделить множество объектов X (множество каротажных данных с месторождения) на M непересекающихся классов из Y (различные типы обстановок осадконакопления). Таким образом, определены понятия пространства

объектов и пространства классов. Как известно, в задачах машинного обучения выделяют два этапа — этап обучения и этап применения. В данном случае обучающая выборка представляет собой набор проинтерпретированных геологом каротажных кривых, где для каждого элемента известно, к какой фациальной обстановке он относится, причем в ней должен присутствовать хотя бы один элемент из каждого класса Y. Неотъемлемым подготовительным этапом для работы алгоритма классификации является также отбор признаков объектов, о котором речь пойдет далее.

Метод решения

Спектральный метод. В задаче распознавания фациальной обстановки спектральный метод представления геофизических полей заключается в разложении каротажных данных (далее для удобства будем называть их также функции каротажа) в ряд Фурье. Пусть $f(h)$ — нормированная функция каротажа в некоторой скважине в отбитом маркерами пласте, переведенная линейным преобразованием на интервал $[-\pi;\pi]$, здесь h - вертикальная координата (глубина). Выявление характерной формы каротажных данных осуществим посредством разложения функции $f(h)$ по базису периодических функций, ортогональных на промежутке $[-\pi;\pi]$:

$$f(h) = S_n(h) + \Delta(h), \tag{1}$$

где $\Delta(h)$ — некоторый остаток;

$$S_n(h) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx). \tag{2}$$

Коэффициенты разложения a_k и b_k могут быть найдены следующим образом:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(h)\,dh; \tag{3}$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(h)\cos kh\,dh, k \in [1;n]; \tag{4}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(h)\sin kh\,dh, k \in [1;n]. \tag{5}$$

Итак,

$$f(h) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx) + \Delta(h). \tag{6}$$

При этом порядок разложения n определяется заданной точностью, с которой необходимо приблизить функцию каротажа $f(h)$ суммой гармонических функций.

Выбор информатиных признаков. Далее для решения задачи классификации в рамках методов машинного обучения необходимо задать признаковое описание объектов. Кроме того, оно должно быть выбрано таким образом, чтобы было возможным эффективно дифференцировать объекты из различных классов. В ходе решения задачи

классификации по типам фациальных обстановок в качестве признаков объектов используется совокупность коэффициентов $\{a_n, b_n\}$ разложения функции каротажа в ряд Фурье. Выбор информативных признаков осуществляется посредством определения порядка разложения функции в ряд необходимого для корректного описания каротажных данных.

Из анализа амплитудных спектров различных каротажных данных было получено, что средняя амплитуда гармоник затухает с ростом частоты (рис. 2), т. е. убывает средняя энергия гармоник. Исходя из этого, для восстановления сигнала и в том числе для эффективного снижения размерности задачи предлагается следующая методика: оставить лишь n первых гармоник, а n следует выбрать таким образом, чтобы остаток, состоящий из высокочастотных гармоник, удовлетворял бы требованию, что средняя (по всем скважинам) относительная ошибка восстановления каротажа ε меньше изначально заданной постоянной величины ε_{max} (это есть та точность, которая необходима для корректного восстановления данных). При этом сама относительная ошибка восстановления каротажа ε для одной скважины при некотором порядке разложения функции в ряд n может быть найдена следующим образом:

$$\varepsilon = \frac{\int_{-\pi}^{\pi} \Delta^2 dh}{\int_{-\pi}^{\pi} f(h)^2 dh} = \frac{\int_{-\pi}^{\pi} \left(f(h) - S_n(h)\right)^2 dh}{\int_{-\pi}^{\pi} f(h)^2 dh}. \tag{7}$$

Поэтому необходимое для спектрального разложения (6) количество коэффициентов определяется так: сначала для каждого примера данных из обучающей выборки считается относительная ошибка восстановления каротажа ε при некотором первоначальном порядке разложения n. При этом, следует заметить, что основную роль играет наличие каротажных данных со значительными неоднородностями — для их корректного восстановления потребуется больше коэффициентов. Далее берется усредненное по всем элементам обучающей выборки значение этой ошибки, и если оно превышает максимально допустимое значение $\varepsilon_{\max}$, то количество коэффициентов увеличивается и рассчитывается относительная ошибка при новом числе коэффициентов. Так происходит до тех пор, пока средняя относительная ошибка превышает $\varepsilon_{\max}$, т. е. пока не достигнута необходимая точность восстановления каротажных данных.

Алгоритм машинного обучения. Следующим этапом в восстановлении обстановки осадконакопления является разбиение множества каротажных данных на классы различных обстановок осадконакопления (или по-другому — фациальных обстановок). Для классификации применяется алгоритм взвешенных ближайших соседей [8], который основан на гипотезе о том, что схожие по признакам объекты принадлежат, как правило, одному классу. Для простоты задачи естественным образом вводится евклидова метрика в пространстве объектов: для двух объектов u и v, которые характеризуются совокупностью признаков $u = (u_1, \ldots, u_n)$ и $v = (v_1, \ldots, v_n)$ (в нашем случае это коэффициенты разложения в ряд Фурье, и их количество при этом считается заданным, оно было определено ранее), расстояние $\rho(u, v)$ между двумя этими объектами в признаковом пространстве будет

$$\rho(u, v) = \sqrt{\sum_{j=1}^{n} (u_j - v_j)^2}. \tag{8}$$

Чем меньше расстояние $\rho(u, v)$, тем более схожи объекты u и v.

Алгоритм взвешенных ближайших соседей заключается в том, что произвольному объекту $u \in X$ из тестовой выборки будет поставлен в соответствие класс $y \in Y$, который име-

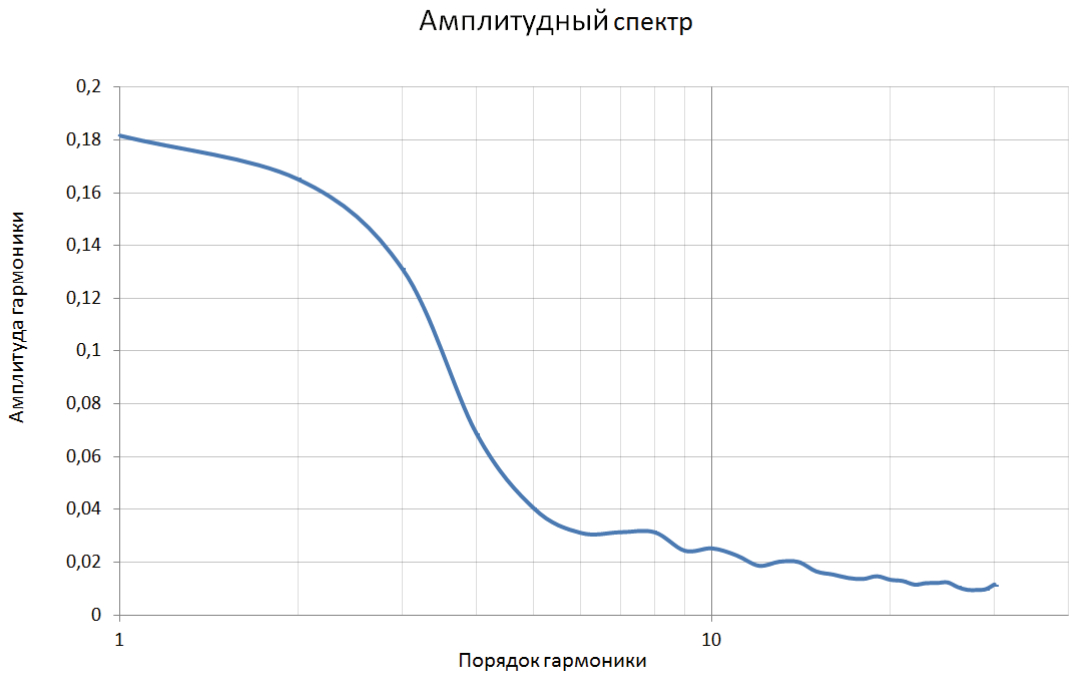


Рис. 2. Амплитудный спектр для примера каротажных данных

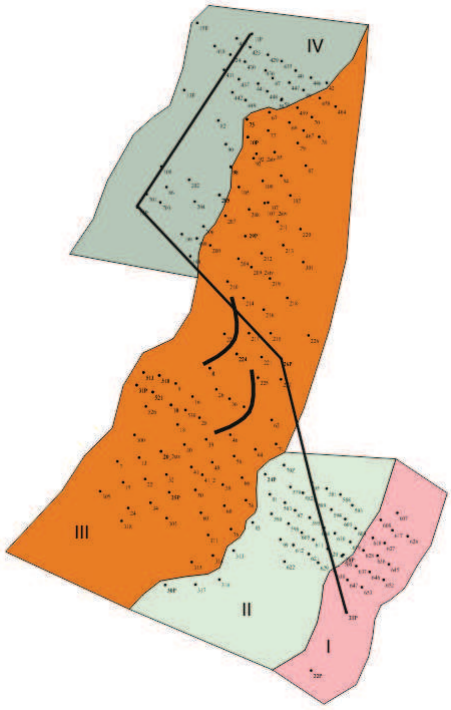
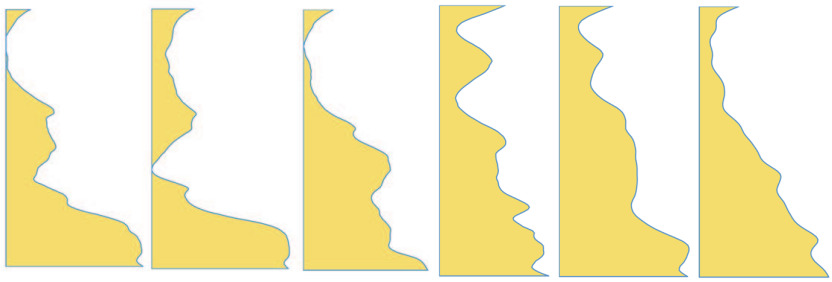


Рис. 3. Расположение скважин месторождения: различные цвета соответствуют различным выделенным фациальным зонам



(а) фация 1 (б) фация 2

(в) фация 3 (г) фация 4

Рис. 4. Примеры данных в обучающей выборке (проинтерпретированные фации)

ет наибольший вес среди его k ближайших соседей из обучающей выборки $X_l = (x_i, y_i)_{i=1}^{l}$. При этом соседство определяется на основе введенной метрики ρ (8) и каждому соседу x_i объекта u присваивается соответствующий вес $w(i, u)$. Имея в распоряжении метрику пространства ρ, введем веса соседей обратно пропорционально квадрату расстояния до объекта u:

$$w(i, u) = \frac{1}{\rho^2(u, x_i)}\,. \tag{9}$$

Оптимальное значение параметра k, числа соседей, обычно определяется про критерию скользящего контроля с исключением объектов по одному. В ходе этой процедуры для каждого объекта $x_i \in X_l$ проверяется, правильно ли он классифицируется по своим k ближайшим соседям. Оптимальным выбирается то значение числа соседей, при котором количество ошибочных классификаций минимально. Таким образом, согласно данному алгоритму для каждого объекта каротажных данных будет определен его класс — тип обстановки осадконакопления.

Эксперимент

Описанная выше методика была применена к восстановлению обстановки осадконакопления на месторождении, которое включало 183 скважины (рис. 3). На рис. 3 различными цветами представлено истинное разбиение по фациальным зонам, которое будет использоваться в конце для анализа ошибки предложенного алгоритма. Входными данными для задачи были каротажные диаграммы, отметки кровли и подошвы пласта по каждой скважине, координаты расположения скважин на плоскости и необходимо было классифицировать все скважины по 4 различным фациальным обстановкам. Для применения алгоритма ближайших соседей была составлена обучающая выборка (рис. 4) с количеством скважин $n_L = 57$, что составляет примерно 30% от их полного числа. На рис. 4

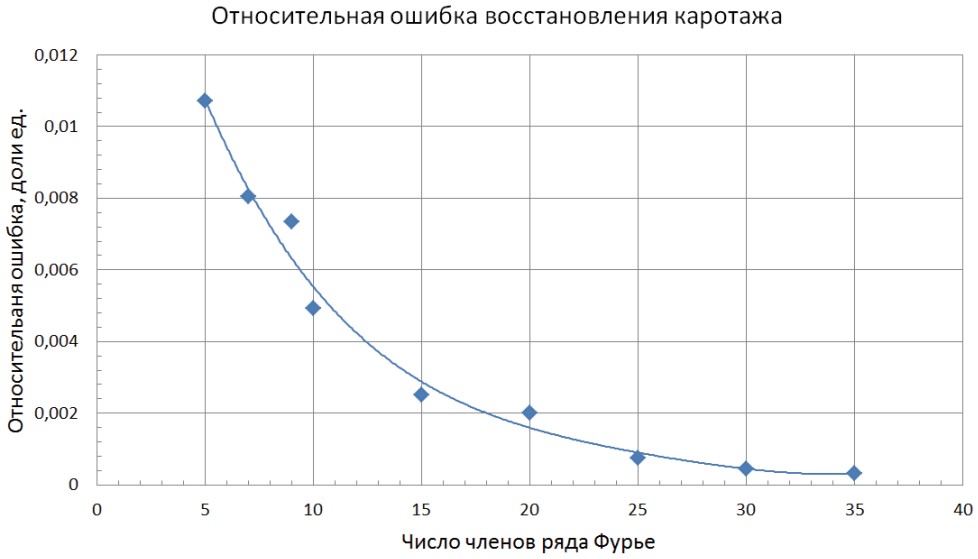


Рис. 5. Пример зависимости относительной ошибки восстановления каротажных данных ε от количества членов разложения для одной скважины

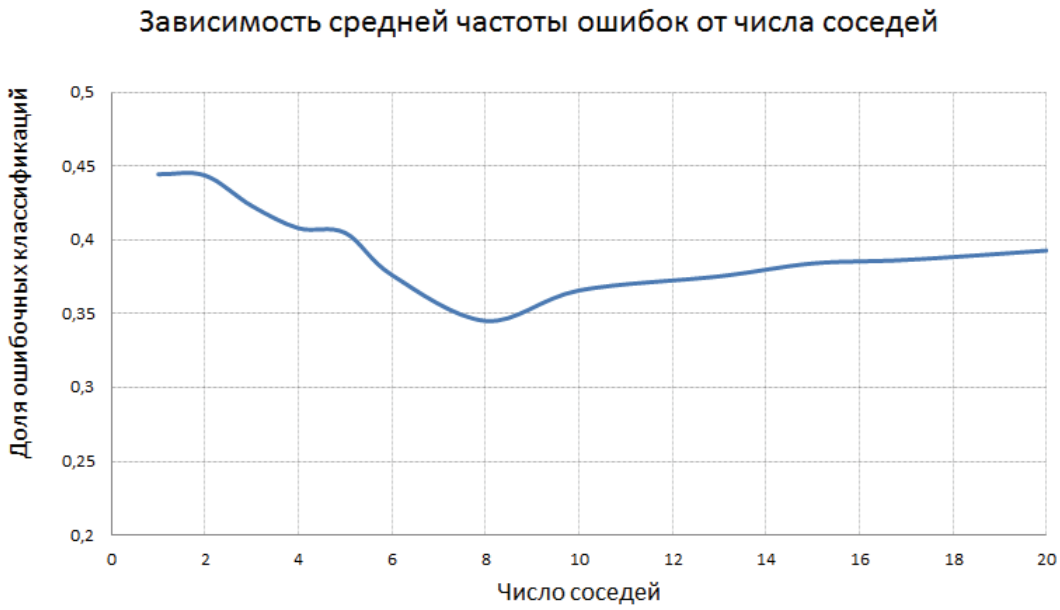


Рис. 6. Зависимость средней частоты ошибок на контрольной выборке от числа соседей, используемых в алгоритме классификации

для каждой скважины представлены следующие данные: по вертикальной оси отложена глубина, по горизонтальной — соответствующее значение каротажа на этой глубине. Для проверки точности работы алгоритма использовалась контрольная выборка, состоящая из всех оставшихся скважин, не вошедших в обучающую выборку.

Вначале был проведен анализ входных данных и определен необходимый порядок разложения функций каротажа в ряд Фурье. По формуле (7) был проведен расчет относительной погрешности восстановления каротажа для каждой скважины. В результате было получено, что при порядке разложения $n = 12$ средняя по всем скважинам ошибка восстановления каротажа $\varepsilon < 0{,}02$ доли ед., и в дальнейших расчетах использовался именно этот порядок разложения каротажных данных в ряд Фурье. Пример зависимости относительной ошибки восстановления каротажных данных для одной скважины от количества членов ряда Фурье показан на рис. 5.

Далее для классификации скважин был применен алгоритм взвешенных ближайших соседей, описанный подробно в предыдущем разделе. При этом оптимальное значение числа соседей было определено по критерию скользящего контроля следующим образом: для различных значений числа соседей вычислялась частота ошибочных классификаций алгоритма на контрольной выборке, которая затем была усреднена по различным наборам обучающих выборок одинаковой длины. Результаты приведены на рис. 6, из которого следует, что оптимальное число соседей для работы алгоритма на этих данных равно $n_{opt} = 8$. При этом в результате работы алгоритма была получена следующая классификация скважин, представленная на на рис. 7, где различные цвета соответствуют различным выделенным обстановкам осадконакопления. Полученное таким образом площадное распределение выделенных фаций в целом, как видно из сравнения рис. 3 и рис. 7, согласуется с морфологией песчаных тел на выбранном участке. Средняя точность алгоритма на контрольных данных при этом составила около 65%.

Заключение

В статье приведена методика распознавания фациальной обстановки, которая была успешно применена к данным реального месторождения. На основе обучающей выборки в результате работы алгоритма машинного обучения были выделены фациальные зоны, при этом процент правильных классификаций на контрольной выборке составил в среднем 65%. В дальнейшем для повышения эффективности распознавания на месторождениях с более сложной геологической структурой планируется реализация более мощных алгоритмов машинного обучения.

Литература

[1] *Муромцев В. С.* Электрометрическая геология песчаных тел — литологических ловушек нефти и газа. М: Недра, 1984. 260 с.

[2] *Matheron G.* The theory of regionalized variables and its applications. Fontainebelau: Center of Geostatistics, 1971. 212 p.

[3] *Matheron G.* The intrinsic random functions and their applications // *Adv. Appl. Probability*, 1973. Vol. 5. P. 439–468.

[4] *Дюбрюль О.* Использование геостатистики для включения в геологическую модель данных. European Association of Geoscientists and Engineers, 2002. 295 с.

[5] *Байков В. А., Бакиров Н. К., Яковлев А. А.* Новые подходы в теории геостатистического моделирования // *Вестник УГАТУ*, 2010. Т. 37, № 2. С. 209–215.

[6] *Бочков А. С., Мухамадеев Д. С.* Новые подходы геостохастического моделирования месторождений // *Доклад ООО «РН-УфаНИПИнефть»*, 2011.

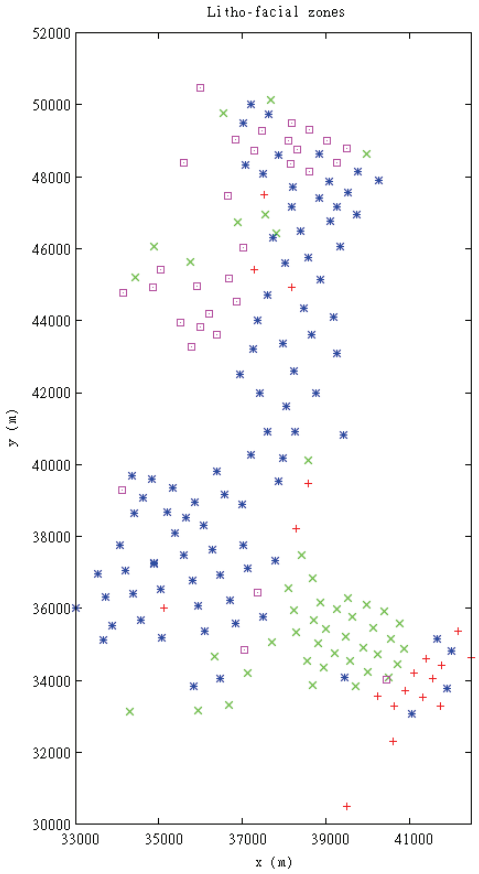


Рис. 7. Площадное распределение фаций

[7] *Байков В. А., Бочков А. С., Яковлев А. А.* Учет неоднородности при геолого-гидродинамическом моделировании Приобского месторождения. Нефтяное хозяйство, 2011.

[8] *Cover T., Hart P.* Nearest neighbor pattern classification // *IEEE Trans. Inform. Theory*, 1967. Vol. 13, No. 1. P. 21–27.

References

[1] *Muromtsev, V. S.* 1984. Electrometric geology of sand bodies — lithological traps of oil and gas. Moscow: Nedra. 260 p. (in Russ.)

[2] *Matheron, G.* 1971. The theory of regionalized variables and its applications. Fontainebelau: Center of Geostatistics. 212 p.

[3] *Matheron G.* 1973. The intrinsic random functions and their applications. *Adv. Appl. Probability* 5:439–468.

[4] *Dubrule, O.* 2002. Application of geostatistics for geological data model. European Association of Geoscientists and Engineers. 295 p. (in Russ.)

[5] *Bajkov, V. A., Bakirov N. K., Yakovlev A. A.* 2010. New methods in the theory of geostatistical modelling. *Vestnik USATU* 37(2):209–215. (in Russ.)

[6] *Bochkov, A. S., Mukhamadeev D. S.* 2011. New approaches to geostochastic modelling of oil fields. *Report RN-UfaNIPIneft LLC.* (in Russ.)

[7] *Bajkov V. A., Bochkov A. S., Yakovlev A. A.* 2011. Accounting of nonhomogeneity in Priobskoye field geological modelling and simulation. Oil industry. (in Russ.)

[8] *Cover T., Hart P.* 1967. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13(1):21–27.