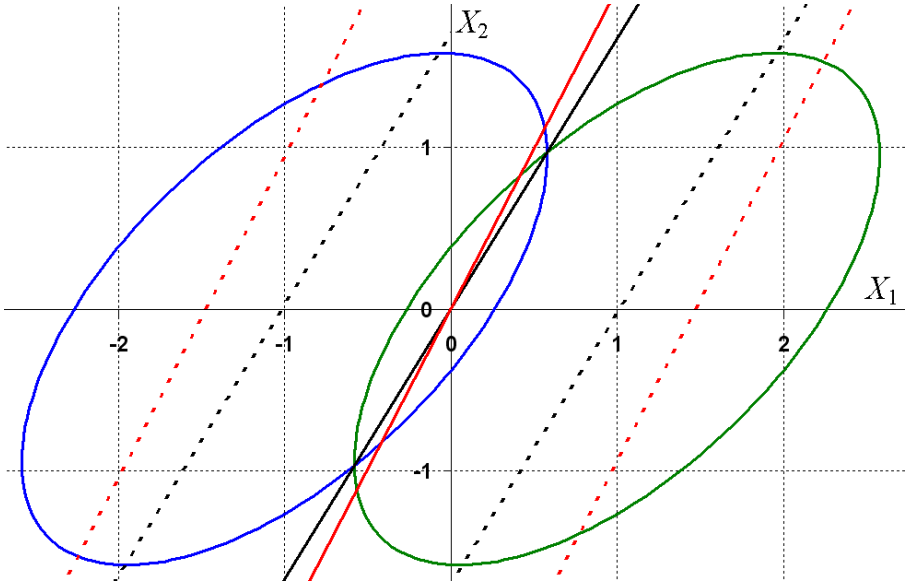
ISSN 2223-3792

Машинное обучение и анализ данных

2016 год **Том 2, номер 3**



Машинное обучение и анализ данных

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретических основ информатики и 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. Журнал зарегистрирован в системе Crossref, doi `http://dx.doi.org/10.21469/22233792`.

- Новостной сайт http://jmlda.org/
- Электронная система подачи статей http://jmlda.org/papers/
- Правила подготовки статей http://jmlda.org/papers/doc/authors-guide.pdf

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

Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2016

Journal of Machine Learning and Data Analysis

The journal Machine Learning and Data Analysis publishes original research papers and reviews of the developments in the field of artificial intelligence, theoretical computer science and its applications. The journal aims to promote the theory of machine learning and data mining and methods of conducting computational experiments. Papers are accepted in English and Russian.

The journal is included in the Russian science citation index RSCI. Information about citation to articles can be found at the Russian science citation index website. ISSN 2223-3792. Mass media registration certificate ЭЛ № ФС 77-55486. The Crossref journal doi is http://dx.doi.org/10.21469/22233792.

- Journal news and archive http://jmlda.org/
- Open journal system for papers submission http://jmlda.org/papers/
- Style guide for authors http://jmlda.org/papers/doc/authors-guide.pdf

The scope of the journal:
— classification, clustering, regression analysis;
— multidimensional statistical analysis;
— Bayesian methods for regression and classification;
— model selection and complexity;
— deep learning;
— Statistical Learning Theory;
— time series forecasting techniques;
— methods of signal processing and speech recognition;
— optimization methods for solving machine learning and data mining problems;
— methods of big data analysis;
— data visualization techniques;
— methods of image processing and recognition;
— text analysis, text mining and information retrieval;
— applied data analysis problems.

Dorodnicyn Computing Centre FRC CSC RAS
Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics
Division "Intelligent Systems"

Moscow, 2016

Содержание

Contents

Комплексирование данных из разнородных источников в задачах моделирования транспортных потоков*

***В. М. Старожилец*[1,2], *Ю. В. Чехович*[1,2]**
starvsevol@gmail.com; chehovich@forecsys.ru
[1]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, д. 44/2
[2]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., д. 9

Исследуется задача агрегации данных с GPS-треков и дорожных датчиков для построения и решения разностной схемы, соотвествующей выбранной математической модели транспортного потока. Отдельно рассматриваются ситуации транспортного потока на самой автомагистрали и потока на въездах и съездах. Для решения обеих задач предложены алгоритмы, а также проведены эксперименты на реальных данных с использованием этих алгоритмов. Для проведения вычислительных экспериментов использованы анонимные трековые данные от сервиса Яндекс.Пробки и данные с дорожных датчиков Центра организации дорожного движения. В качестве автомагистрали рассматривалась Московская кольцевая автомобильная дорога.

Ключевые слова: *восстановление данных; МКАД; данные GPS-треков; дорожные датчики*

DOI: 10.21469/22233792.2.3.01

1 Введение

Работа посвящена проблеме агрегации данных из разных источников, используемых в задаче моделирования транспортных потоков [1]. Для моделирования транспортных потоков используются данные о скорости и числе проехавших по рассматриваемому участку автодороги автотранспортных средств (АТС). Эти данные могут быть получены с помощью *GPS-треков* (данные трекового типа) и *дорожных датчиков*, которые имеют различные свойства. Дорожные датчики измеряют скорость и число АТС с приемлемой точностью, погрешность измерений не превышает 10% в зависимости от типа датчика [2], но не всегда покрывают транспортную сеть на требуемом уровне. В то же время данные с GPS-треков (трекового типа) имеют недостаточную точность, так как фиксируют небольшой процент[1] от общего числа проехавших АТС, однако полностью покрывают транспортную сеть. Поэтому основная идея работы заключается в том, чтобы агрегировать данные с дорожных датчиков и GPS-треков для получения более точных данных с большим покрытием транспортной сети.

Моделирование транспортных потоков основано на их сходстве с жидкой или газовой средой. В частности, базовая модель Лайтхилла–Уизема–Ричардса (LWR) [3–5] основана на предположении о существовании взаимно-однозначной зависимости между скоростью и плотностью потока АТС и сохранении числа АТС в транспортной сети. В современном макроскопическом подходе транспортный поток описывается нелинейной системой

*Работа выполнена при частичной финансовой поддержке РФФИ проект № 14-07-00685

[1]Рассчитать долю трековых АТС в общем потоке можно, взяв отношение числа трековых АТС, к числу АТС зафиксированному датчиком для детектора и сегмента под ним. Рассчитанные по данной методике на имеющихся у авторов данных значения имеют большой разброс.

гиперболических дифференциальных уравнений в частных производных второго порядка в различных постановках [6–13]. Такая система требует точных входных данных с хорошим покрытием транспортной сети для выбранной модели. Эти требования к данным необходимы для построения и разрешения разностной схемы, соотвествующей выбранной модели, и заданию граничных условий. Ранее было показано [14], что абсолютно точных данных без пропусков для задач транспортного типа не существует.

Ранее задача агрегирования произвольных данных рассматривалась в работах [15–18]. Предложенные там методы не учитывают специфику задачи агрегации данных для моделирования транспортных потоков. Для решения проблемы получения точных данных с хорошим покрытием транспортной сети в работе предлагается метод агрегирования данных GPS-треков и дорожных датчиков. Однако GPS-треки на автомагистрали и на въездах и съездах с нее сильно отличаются из-за существенно меньшего потока АТС на въездах и съездах по сравнению с автомагистралью, поэтому данные с въездов и съездов рассматриваются отдельно от данных с автомагистрали. В частности, для данных с автомагистрали предложен метод агрегации данных GPS-треков и дорожных датчиков, основанный на построении линейной модели для скорости и числа АТС. Критерием качества полученной модели является среднеквадратичная ошибка между оцененным числом проехавших АТС и реальным, а также коэффициент корреляции между ними. Число реально проехавших АТС определяется по данным дорожных датчиков. Для данных на въездах и съездах был разработан метод восстановления суммарного потока на основе сохранения числа АТС в транспортной сети с использованием оценки числа проехавших АТС, полученной с помощью модели на данных с автомагистрали. В работе был проведен вычислительный эксперимент на данных дорожных датчиков и GPS-треков для Московской кольцевой автомобильной дороги за 2012 г.

2 Постановка задачи

Поскольку в работе рассматривается задача агрегации даных для двух принципиально различных дорожных конфигураций: автомагистрали и въезд–съезд, то необходимо поставить задачу для каждой из этих конфигураций. Поставленные задачи будут отличаться подходом к агрегации данных и способом проверки полученного решения.

2.1 Задача агрегации для данных с автомагистрали

Пусть $N_{\text{track}} \in \mathbb{N}$ и $V_{\text{track}} \in \mathbb{R}_+$ — число АТС и их средняя скорость, полученные из данных GPS-треков для определенного участка дороги. Обозначим через $N_{\text{det}} \in \mathbb{N}$ и $V_{\text{det}} \in \mathbb{R}_+$ число АТС и их среднюю скорость, полученные с помощью дорожных датчиков для того же участка дороги.

Необходимо найти функцию $f : \mathbb{N}\times\mathbb{R}_+\times\mathbb{R}^m \to \mathbb{R}_+$ от $N_{\text{track}}\times V_{\text{track}}$ и вектора параметров $\mathbf{a} \in R^m$, где m — сложность модели, которая аппроксимирует число АТС, проехавших по сегменту автомагистрали

$$N_{\text{est}} = f(\mathbf{a}|N_{\text{track},i}, V_{\text{track},i})$$

и является решением следующей задачи:

$$\sigma(\mathbf{a}|\mathbf{N}_{\text{track}}, \mathbf{V}_{\text{track}}, \mathbf{N}_{\text{det}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{a}|N_{\text{track},i}, V_{\text{track},i}) - N_{\text{det},i})^2} \to \min_{\mathbf{a}}, \tag{1}$$

где $\mathbf{N}_{\text{track}} = [N_{\text{track},i}] \in \mathbb{N}^n$, $\mathbf{V}_{\text{track}} = [V_{\text{track},i}] \in \mathbb{R}_+^n$ и $\mathbf{N}_{\text{det}} = [N_{\text{det},i}] \in \mathbb{N}^n$ — вектора значений $N_{\text{track}}, V_{\text{track}}$ и N_{det} в момент времени i, а n — число двухминутных интервалов

в выбранном временно́м промежутке. Предполагается, что f зависит только от трековых переменных $\mathbf{N}_{\text{track}}$ и $\mathbf{V}_{\text{track}}$, что позволяет ее использовать на участках автомагистрали, которые не покрываются дорожными датчиками.

Однако скорость V_{track} вычисляется по малой доле АТС, поэтому она может сильно отличаться от реальной скорости потока АТС. Для уменьшения влияния этого отличия предлагается вместо оригинальных значений V_{track} использовать модифицированные значения V_{est}, определяемые выражением:

$$V_{\text{est}} = b_1 + b_2 V_{\text{track}},$$

где коэффициенты b_1 и b_2 являются решением следующей задачи:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_1 + b_2 V_{\text{track},i} - V_{\text{det},i})^2} \to \min_{b_1,b_2}, \tag{2}$$

а $V_{\text{track},i} \in \mathbb{R}_+$ и $V_{\text{det},i} \in \mathbb{R}_+$ — значения V_{track} и V_{det} в момент времени i. Таким образом, задача (1) преобразуется в

$$\sigma(\mathbf{a}|\mathbf{N}_{\text{track}}, \mathbf{V}_{\text{est}}, \mathbf{N}_{\text{det}}) \to \min_{\mathbf{a}}. \tag{3}$$

Задача (3) решается на сегментах, для которых известны как данные с GPS-треков, так и данные с дорожных датчиков. Это множество сегментов разбивается на обучающую и контрольные выборки. Способ разбиения изложен в подразд. 5.2.

Внешним критерием качества полученного решения является коэффициент корреляции $\text{corr}(\mathbf{N}_{\text{est}}, \mathbf{N}_{\text{det}})$, где $\mathbf{N}_{\text{est}} = [N_{\text{est},i}] \in \mathbb{R}_+^n$ и $\mathbf{N}_{\text{det}} = [N_{\text{det},i}] \in \mathbb{N}^n$ — вектора значений N_{est} и N_{det} в момент времени i.

2.2 Задача агрегации для данных со въездов–съездов

Въезды и съезды образуют перекрестки с автомагистралью. Перекресток — это место пересечения, примыкания или разветвления дорог на одном уровне, ограниченное воображаемыми линиями, соединяющими соответственно противоположные, наиболее удаленные от центра перекрестка начала закруглений проезжих частей.

Для оценки суммарного количества АТС, въехавших и съехавших с автомагистрали, используется уравнение баланса, заключающееся в том, что на въездах и съездах число въезжающих АТС должно равняться числу выезжающих:

$$N_{\text{ain}} + N_{\text{in}} = N_{\text{aout}} + N_{\text{out}},$$

где $N_{\text{ain}} \in \mathbb{R}_+$ и $N_{\text{aout}} \in \mathbb{R}_+$ — оценка числа въехавших на перекресток по автомагистрали и выехавших по автомагистрали после перекрестка АТС, полученная после решения задачи (3); N_{in} — суммарное число въехавших по въездам АТС; N_{out} — суммарное число съехавших по съездам АТС.

Рассмотрим, как вычислить значение N_{in}; значение N_{out} вычисляется аналогично. По определению $N_{\text{in}} = \sum_{k \in K_{\text{in}}} N_{\text{det},k}$, где $K_{\text{in}} = \{1, \ldots, K\}$ — множество индексов съездов, а $N_{\text{det},k}$ — значение N_{det} на k-м въезде. Но не для всех $k \in K_{\text{in}}$ значение $N_{\text{det},k}$ известно. Поэтому рассмотрим разбиение множества K_{in} на два непересекающихся подмножества K_{indet} и K_{intrack}: $K_{\text{in}} = K_{\text{indet}} \cup K_{\text{intrack}}$ и $K_{\text{intrack}} \cap K_{\text{indet}} = \varnothing$. Множество K_{indet} состоит из индексов въездов, для которых известно N_{det}, а множество K_{intrack} состоит из индексов въездов, для которых неизвестно N_{det}.

Чтобы получить оценку N_{det} для въездов из K_{intrack}, предлагается использовать подход описанный в подразд. 2.1:

$$\sigma(\mathbf{a}|\mathbf{N}_{\text{track}}, \mathbf{V}_{\text{est}}, \mathbf{N}_{\text{det}}) \to \min_{\mathbf{a}},$$

где $\mathbf{N}_{\text{track}} = [N_{\text{track},i}] \in \mathbb{N}^n$, $\mathbf{V}_{\text{track}} = [V_{\text{track},i}] \in \mathbb{R}_+^n$ и $\mathbf{N}_{\text{det}} = [N_{\text{det},i}] \in \mathbb{N}^n$ — вектора значений $N_{\text{track}}, V_{\text{track}}$ и N_{det} для въезда из множества K_{indet} в момент времени $i \in I_{\text{in}}$, $n = |I_{\text{in}}|$. Множество I_{in} — это множество индексов временны́х интервалов, таких что в момент времени $i \in I_{\text{in}}$ известно число въехавших АТС $N_{\text{in},k}$ для всех въездов $k \in K_{\text{in}}$. Для получение необходимых данных на съездах нужно решить аналогичную задачу. Введем суммарное число въехавших по въездам АТС

$$N_{\text{in}} = \sum_{k \in K_{\text{indet}}} N_{\text{det},k} + \sum_{k' \in K_{\text{intrack}}} N_{\text{est},k'},$$

где $N_{\text{det},k}$ — значение N_{det} на k-м въезде; $N_{\text{est},k'}$ — значение N_{est} на k'-м въезде. Аналогично определяется суммарное число съехавших по съездам АТС N_{out}.

Требуется построить алгоритм нахождения таких значений N_{estin} и N_{estout}, чтобы они удовлетворяли уравнению баланса и различие между ними и значениями N_{in} и N_{out} в моменты времени из множества I_{in} и I_{out} соответсвенно было не слишком велико, что формализуется следующим образом:

$$\left.\begin{aligned} (N_{\text{ain}} + N_{\text{estin}} - N_{\text{aout}} - N_{\text{estout}})^2 &\to \min_{N_{\text{estin}}, N_{\text{estout}}}; \\ \text{s.t. } \frac{1}{n}\left(\sum_{i \in I_{\text{in}}} |N^i_{\text{estin}} - N^i_{\text{in}}| + \sum_{i' \in I_{\text{out}}} |N^{i'}_{\text{estout}} - N^{i'}_{\text{out}}|\right) &< \delta\,, \end{aligned}\right\} \tag{4}$$

где δ — допустимое отличие оценки числа АТС на съездах и въездах $N^{i'}_{\text{estout}}$ и N^i_{estin} от наблюдения $N^{i'}_{\text{out}}$ и N^i_{in}; I_{out} — множество аналогичное I_{in} для съездов; N^i_{in} и N^i_{estin} — значения N_{in} и N_{estin} в момент времени i; $N^{i'}_{\text{out}}$ и $N^{i'}_{\text{estout}}$ — значения N_{out} и N_{estout} в момент времени i'; $n = |I_{\text{in}}| + |I_{\text{out}}|$.

3 Выбор модели для предсказания числа автотранспортных средств

Для решения задачи (3) необходимо задать вид функции f. Было показано, что f линейно зависит от числа трековых АТС N_{track}, скорости V_{track} и оценки плотности потока [19]. Для проверки наличия дополнительных зависимостей была построена скрипичная диаграмма [20], изображенная на рис. 1. Рисунок 1 показывает, что линейной аппроксимации недостаточно для приближения реального числа АТС. Чтобы учесть найденную нелинейность, предлагается добавить слагаемое $\log(N_{\text{track}})$, так как средние значения на рис. 1 лежат на кривой, похожей на график логарифма. Таким образом, будем искать функцию f в виде:

$$f(\mathbf{a}|N_{\text{track},i}, V_{\text{est},i}) = a_0 + a_1 N_{\text{track},i} + a_2 \log(N_{\text{track},i}) + a_3 V_{\text{est},i} + a_4 \frac{N_{\text{track},i}}{V_{\text{est},i}}\,. \tag{5}$$

Также кроме модели вида (5) были рассмотрены и другие модели, такие как модель числа трековых АТС:

$$f(\mathbf{a}|N_{\text{track},i}) = a_0 + a_1 N_{\text{track},i}\,, \tag{6}$$

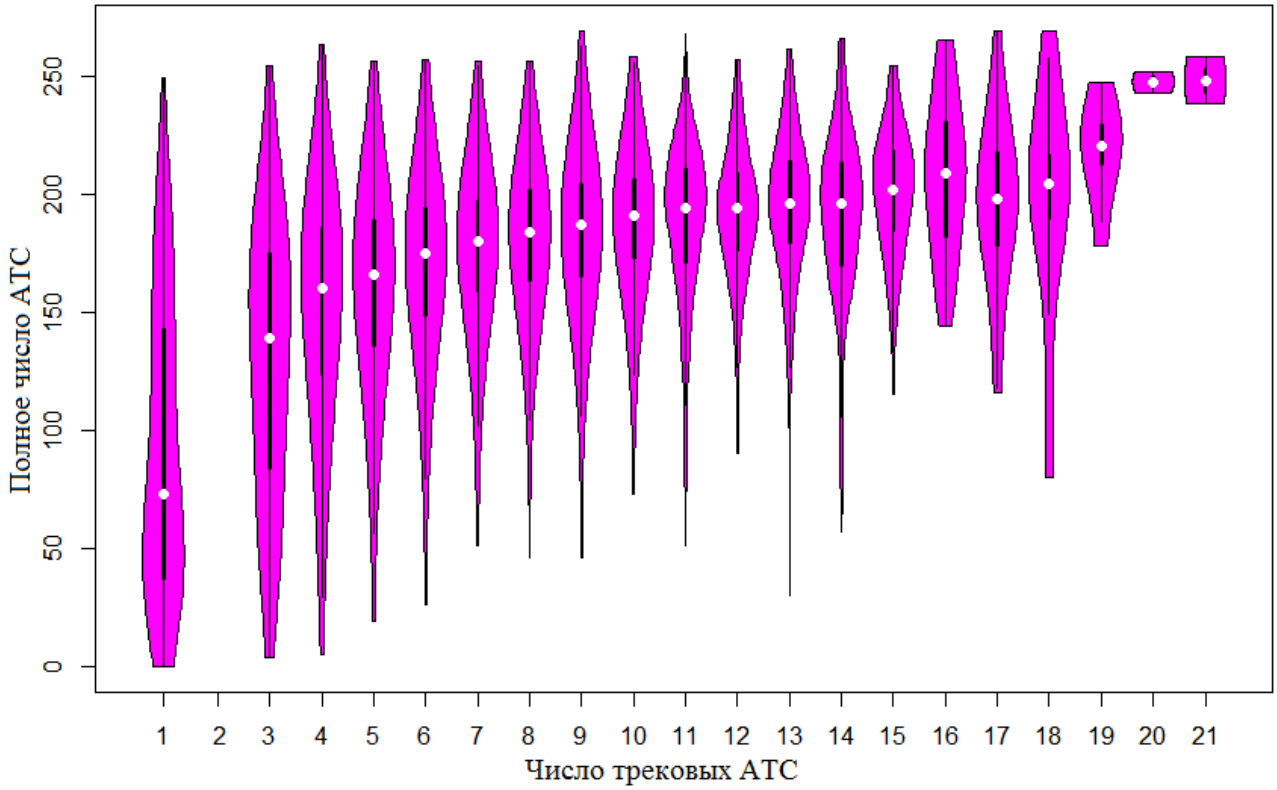


Рис. 1 Скрипичная диаграмма зависимости полного числа АТС от числа трековых АТС

модель числа АТС и скорости

$$f(\mathbf{a}|N_{\text{track},i}, V_{\text{est},i}) = a_0 + a_1 N_{\text{track},i} + a_2 V_{\text{est},i} \tag{7}$$

и модель с логарифмом числа АТС и скоростью

$$f(\mathbf{a}|N_{\text{track},i}, V_{\text{est},i}) = a_0 + a_1 \log\left(N_{\text{track},i}\right) + a_2 V_{\text{est},i}\,. \tag{8}$$

4 Алгоритм восстановления суммарного потока на въездах и съездах

Алгоритм восстановления суммарного потока на въездах и съездах состоит из следующих шагов:

- задать перекресток, на котором будет решаться задача восстановления суммарного потока, и определить сегменты, соответствующие въездам и съездам, а также сегменты с которых берутся данные о N_{ain} и N_{aout};
- определить въезды и съезды, принадлежащие множествам K_{intrack} и K_{outtrack}. Уже имеющееся на них небольшое число данных используем для определения параметров случайного Пуассоновского процесса, соответствующего данному въезду или съезду;
- используя полученные распределения на въездах и съездах из множеств K_{intrack} и K_{outtrack} вместо реальных данных, а также данные с самого МКАД (которые считаем верно восстановленными в подразд. 5.2) и данные с въездов и съездов из множеств K_{indet} и K_{outdet}, решаем задачу (4).

Алгоритм 1 Алгоритм восстановления суммарных значений потока на въездах/съездах, использующий уравнение баланса

если $N \neq 0$ **то**
  если $N > 0$ **то**
    $\max\_extra\_cars = N_{\text{out}} + (\max\_cars\_in - N_{\text{in}})$
    если $N < \max\_extra\_cars$ **то**
      $N_{\text{estout}} = N_{\text{out}} - N \cdot N_{\text{out}} / \max\_extra\_cars$
      $N_{\text{estin}} = N_{\text{in}} + N \cdot (\max\_cars\_in - N_{\text{in}}) / \max\_extra\_cars$
    иначе
      $N_{\text{estout}} = 0$
      $N_{\text{estin}} = \max\_cars\_in$
  иначе
    $\max\_extra\_cars = N_{\text{in}} + (\max\_cars\_out - N_{\text{out}})$
    если $|N| < \max\_extra\_cars$ **то**
      $N_{\text{estin}} = N_{\text{in}} - N \cdot N_{\text{in}} / \max\_extra\_cars$
      $N_{\text{estout}} = N_{\text{out}} + N \cdot (\max\_cars\_out - N_{\text{out}}) / \max\_extra\_cars$
    иначе
      $N_{\text{estin}} = 0$
      $N_{\text{estout}} = \max\_cars\_out$

Для решения поставленной задачи предлагается использовать алгоритм 1, использующий понятие дисбаланса $N = N_{\text{ain}} + N_{\text{in}} - N_{\text{aout}} - N_{\text{out}}$. Также введем следующие обозначения $\max\_cars\_in = k \max\_cars\_per\_enter/exit$ — максимальное число въезжающих по въездам АТС; $\max\_cars\_out = k' \max\_cars\_per\_enter/exit$ — максимальное число съезжающих по съездам АТС, где $\max\_cars\_per\_enter/exit$ — максимальное число АТС, которые могут проехать по съезду/въезду за 2 мин; k — число въездов; k' — число съездов. Значение $\max\_cars\_per\_enter/exit$ равно 60 АТС [21], т. е. 1 АТС за каждые 2 с.

5 Вычислительный эксперимент

В работе проведен вычислительный эксперимент на реальных данных с МКАДа за 2012 г. и проверен предложенный подход к агрегации данных с автомагистралей.

5.1 Описание данных

В данной работе используются анонимные данные с GPS-треков и дорожных датчиков за 2012 год Данные с GPS-треков представляют собой набор сегментов, каждый из которых соотносится с некоторым участком автодороги. Объединение сегментов покрывает весь рассматриваемый участок транспортной сети. Для каждого сегмента известно число проехавших за двухминутный интервал трековых АТС N_{track} и среднее время их проезда по нему, из которого впоследствии рассчитывается скорость V_{track}. Также если за 2 мин по данному сегменту проехало менее трех трековых АТС, то они не учитываются.

Для каждого датчика известно его местоположение на автомагистрали. Данные с датчиков состоят из числа проехавших за 2 мин АТС N_{det} для каждой из полос и их скорости V_{det}. Заметим, что в данных с датчиков также могут быть ошибки, например на рис. 2 показаны данные датчика, в которых отсутствуют записи за 4 ч.

5.2 Эксперимент на автомагистрали

В экспериментах вместо числа АТС использовалась плотность АТС, чтобы учесть различные длины сегментов. Обозначим плотность АТС на участке автомагистрали, полу-

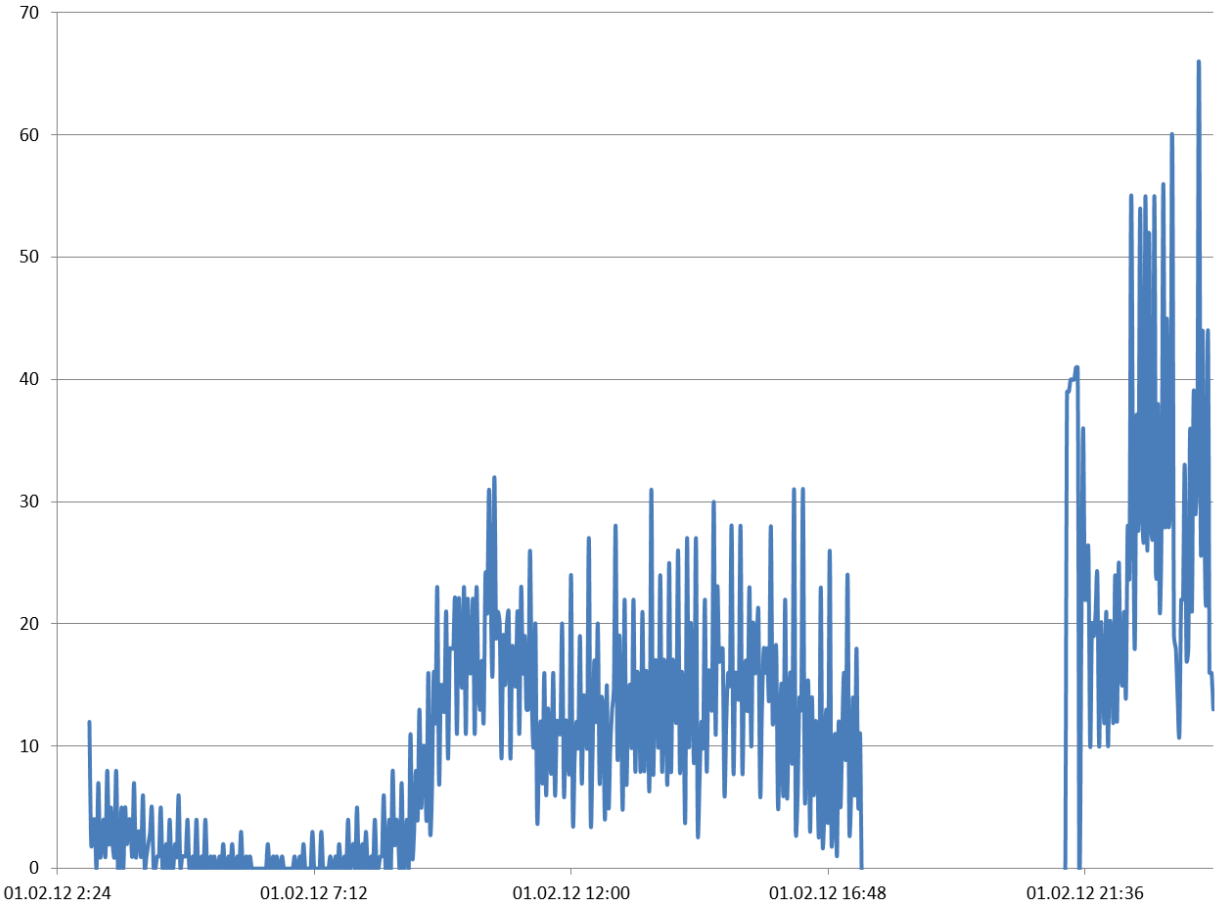


Рис. 2 Число зарегистрированных датчиком АТС в зависимости от времени суток. Детектор перестает фиксировать АТС в 17:25 и начинает в 21:15

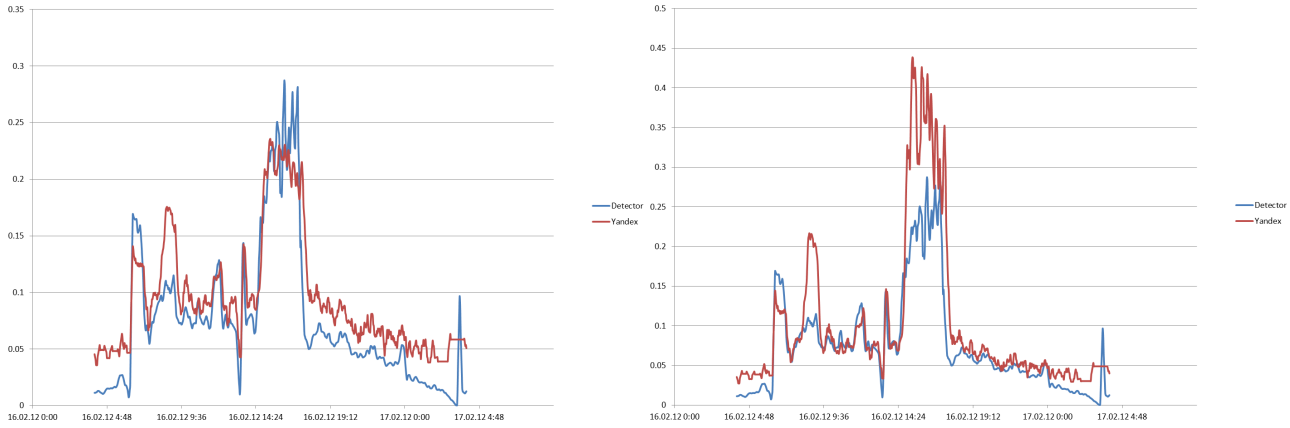
ченную с помощью данных дорожных датчиков $\rho_{\text{det}} = N_{\text{det}}/l$, где l — длина участка автомагистрали в данных GPS-треков. $\boldsymbol{\rho}_{\text{est}} \in \mathbb{R}_+^n$ и $\boldsymbol{\rho}_{\text{det}} \in \mathbb{R}_+^n$ — вектора соответсвующих по времени значений ρ_{est} и ρ_{det}, где n — число двухминутных интервалов в выбранном временно́м промежутке.

Для решения задачи (3) необходимо сначала получить преобразование скорости, решив задачу (2). Решением задачи (2) является следующее выражение:

$$V_{\text{est}} = 12{,}4 + 0{,}639 V_{\text{track}}. \tag{9}$$

На рис. 3 показана зависимость плотности АТС от времени суток в случае использования преобразования (9) (рис. 3, *а*) и в случае использования скоростей, полученных с GPS-треков V_{track} (рис. 3, *б*). При использовании преобразования (9) ошибка аппркосимации на обучении $\sigma(\mathbf{a_1}|\rho_{\text{track}}, \mathbf{V}_{\text{est}}, \rho_{\text{det}}) = 0{,}03$ и корреляция $\text{corr}(\rho_{\text{est}}, \rho_{\text{det}}) = 0{,}787$, в то время как при использовании данных с GPS-треков ошибка аппроксимации $\sigma(\mathbf{a_2}|\rho_{\text{track}}, \mathbf{V}_{\text{track}}, \rho_{\text{det}}) = 0{,}042$ и корреляция $\text{corr}(\rho_{\text{est}}, \rho_{\text{det}}) = 0{,}672$. Это означает, что использование преобразования (9) повышает качество аппроксимации.

Далее рассмотрим метод улучшения качества аппроксимации с помощью построения нескольких моделей для данных с различными значениями плотности АТС, определяемыми по данным датчиков. Для этого возьмем множество данных за февраль L для датчика и подсегмента под ним и выделим из них множество $H \subset L$ — данные, соответствующие



(*а*) (*б*)

Рис. 3 Графики полученных с помощью моделей плотностей в случае с использованием модели для скорости (*а*) и без ее использования (*б*)

плотности АТС более 0,05 АТС/м, которая является переходной между фазами свободного и синхронизированного потока для МКАД [1].

Задача (3) решается для данных L и H отдельно с получением векторов параметров $\mathbf{a}_L$ и $\mathbf{a}_H$ соответственно. Решением задачи (3) на данных H является следующее выражение:

$$N_{\text{est}} = 157{,}78 + 4{,}54N_{\text{track}} - 4{,}59\log(N_{\text{track}}) + 0{,}153V_{\text{est}} - 85{,}069\frac{N_{\text{track}}}{V_{\text{track}}}, \tag{10}$$

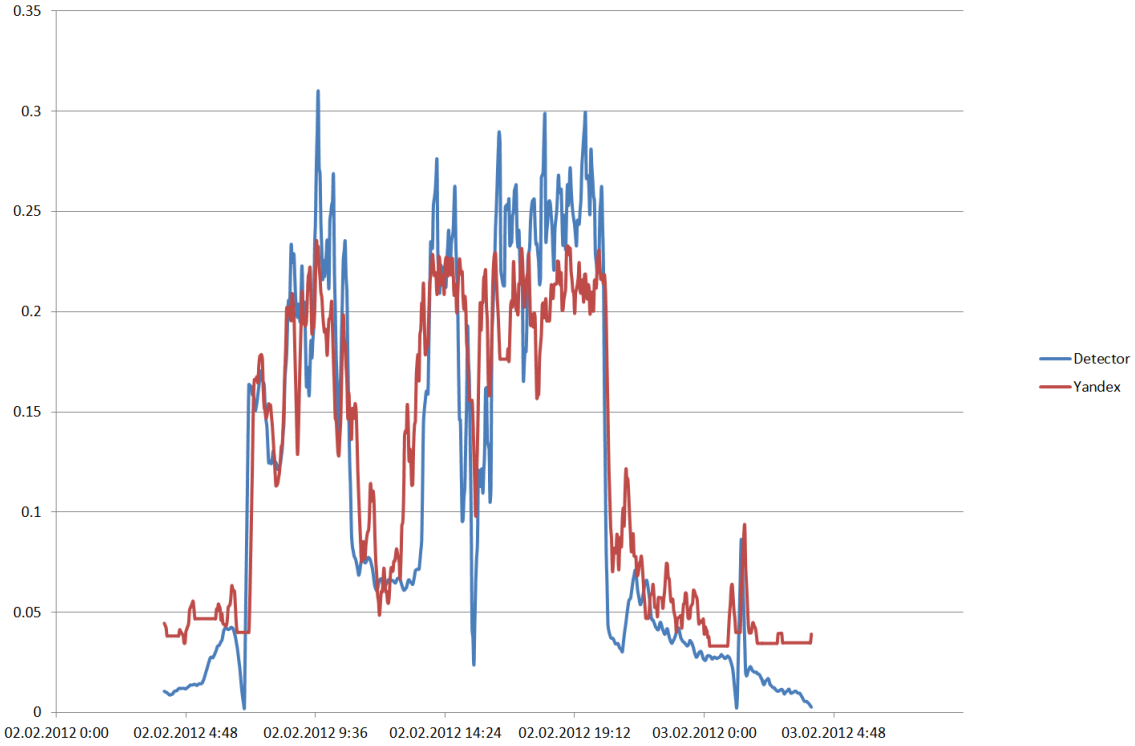
а на данных L:

$$N_{\text{est}} = 117{,}75 + 2{,}11N_{\text{track}} + 41{,}55\log(N_{\text{track}}) - 0{,}327V_{\text{est}} - 128{,}89\frac{N_{\text{track}}}{V_{\text{est}}}. \tag{11}$$

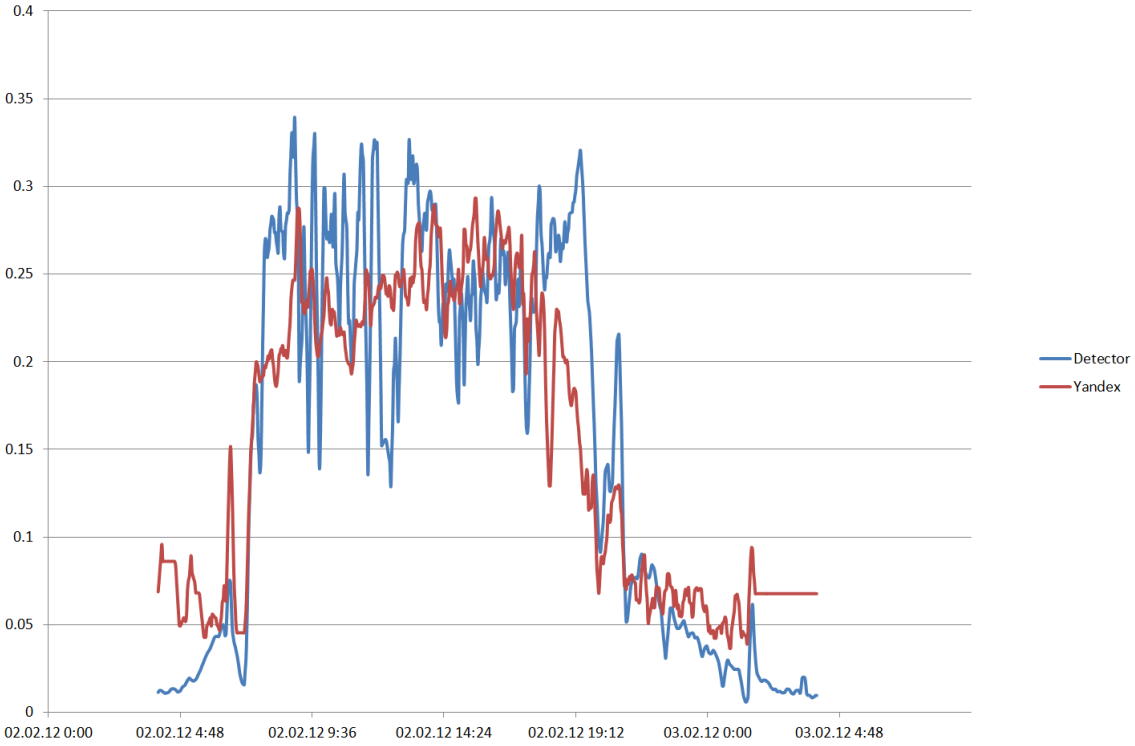
Первая модель использовалась при плотности выше 0,1 АТС/м (плотность, соответствующая началу синхронизированной фазы [1]), вторая — при плотности ниже 0,1 АТС/м. Корреляция на обучении составила 0,787, средняя ошибка — 0,03, а сравнение результата ρ_{det} с ρ_{est} показано на рис. 4, *а*. Для контроля выбраны четыре пары датчик–сегмент и проведены вычисления с использованием моделей (10) и (11). На контроле корреляция составила 0,823, 0,80, 0,85, и 0,65, средняя ошибка — 0,0363, 0,0382, 0,0339 и 0,0393 соответственно. Сравнение результата ρ_{det} с ρ_{est} показано на рис. 4, *б* для одной из тестовых пар.

Также был проведен эксперимент для проверки возможности использования построенной модели для получения оценки количества проехавших АТС в режиме реального времени с использование данных GPS-треков. Чтобы обучить модель в этом случае, необходимо использовать исторические данные за некоторый промежуток времени до дня, для которого надо получить оценку. В рассматриваемом случае для обучения брались данные за каждые 7 дней перед рассматриваемым днем, который является контрольным временным интервалом. На рис. 5 показана заисимость функции ошибки и коэффициента корреляции от дня на усредненных за 10 мин данных для построенных моделях для каждого дня.

На рис. 6 показаны результаты работы моделей (6)–(8). Сравнение рассматриваемых моделей приведено в табл. 1. Из табл. 1 следует, что лучшей является модель (5).

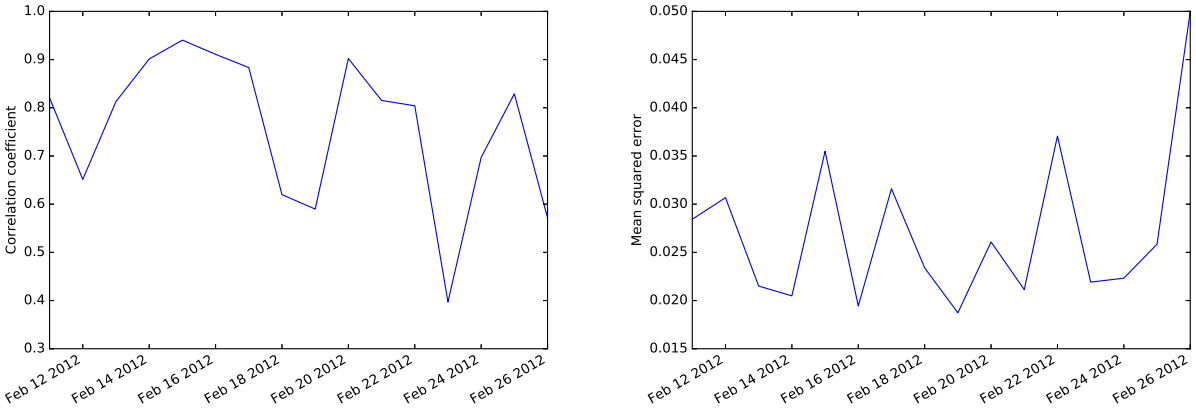


(*а*)



(*б*)

Рис. 4 Средняя за 10 мин плотность АТС, рассчитанная по данным датчика и аппроксимированным трековым данным на обучении (*а*) и на контроле (*б*)



(*а*) (*б*)

Рис. 5 Корреляция для усреднённых за 10 мин данных (*а*) и среднеквадратичная ошибка на контроле для эксперимента с обучением по 7 дням (*б*)

В табл. 2 приведено сравнение среднеквадратической ошибки моделей (5)–(7), (8) при большой плотности $\rho_{\text{det}} > 0{,}2$. Из табл. 2 следует, что модель (5) значительно лучше при больших плотностях, чем модели (6) и (8) и лучше модели (7).

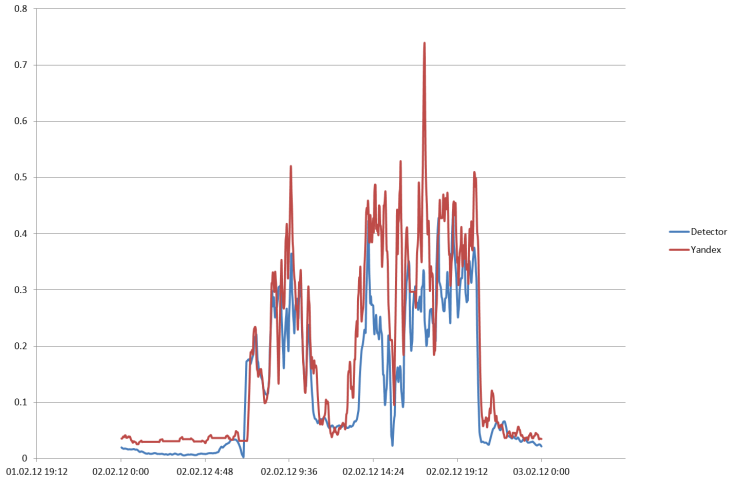
5.3 Эксперимент на въездах и съездах

Приведем на примере двух перекрестков результат восстановления числа въехавших АТС за 02.02.2012. На одном из перекрестков датчиками закрыты все въезды и выезды, кроме одного въезда, результат работы алгоритма 1 представлен на рис. 7. На рис. 7 показано, что результат работы алгоритма 1 (красная кривая) проходит в области, соответствующей сумме данных с датчика и данных с GPS-треков (зеленые точки), т. е. ограничение в задаче (4) выполняется с достаточной точностью.

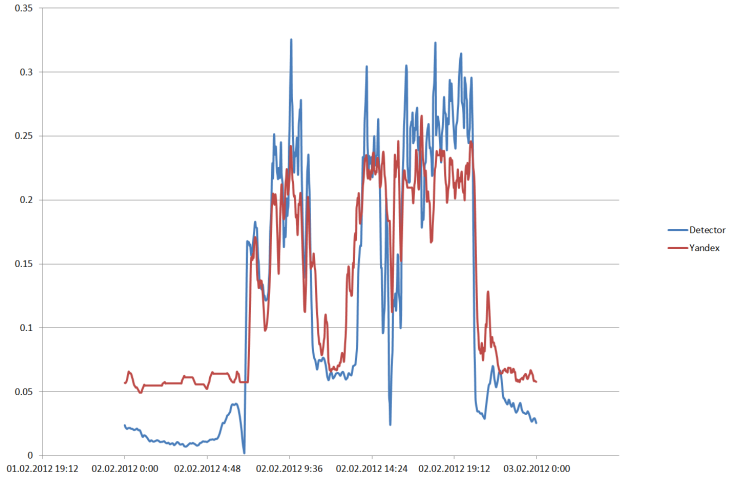
На втором перекрестке датчиками закрыты оба въезда, но один из датчиков фиксировал проехавшие АТС полчаса за день (до 5 АТС), и этот въезд был также включен в множество K_{intrack}, данных же трекого типа на данном въезде нет, результат работы алгоритма 1 представлен на рис. 8. На рис. 8 показано, что результат работы алгоритма 1 (красная кривая) проходит в области, соответствующей сумме данных с датчика и данных с GPS-треков (зеленые точки), т. е. ограничение в задаче (4) выполняется с достаточной точностью. Также на рис. 8 показано, что N_{estin} в большинство двухминутных интервалов отличается от данных закрытого датчиком въезда (синяя кривая) менее чем на 5 АТС, а иногда полностью совпадает с ним. Таким образом, данные с закрытого плохим датчиком въезда подтверждаются (малое число зафиксированных АТС), а также становится понятна причина отсутствия данных трекового типа на данном въезде — слишком слабый поток АТС. Все значения на рис. 7 и 8 усреднены за 10 мин.

6 Обсуждение результатов

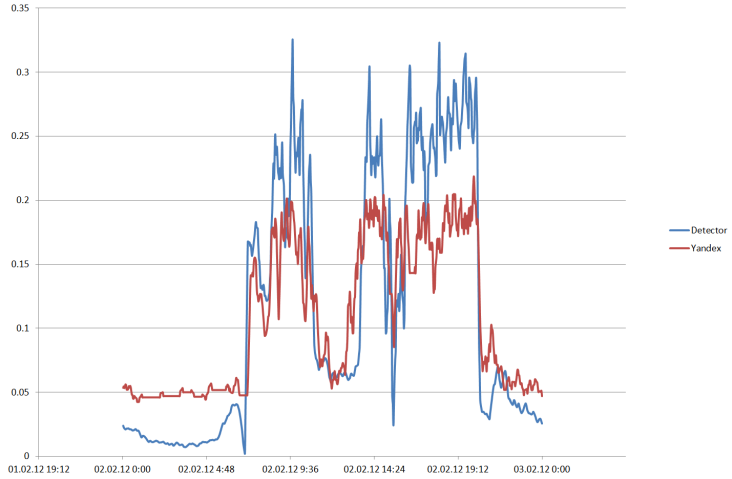
В работе были изложены два алгоритма агрегации данных: на автомагистрали и на въездах и съездах. Для обоих алгоритмов были поставлены эксперименты на реальных данных с целью проверки их работоспособности. Таким образом, получены следующие результаты:



(*а*)



(*б*)



(*в*)

Рис. 6 Плотность АТС для результатов обучения модели (6) (*а*), (7) (*б*) и (8) (*в*)

Таблица 1 Сравнение моделей при всех значениях плотности ρ_{det}

Модель	Среднеквадратическая ошибка	Корредяция
(5)	**0,03**	**0,787**
(6)	0,031	0,781
(7)	0,0326	0,765
(8)	0,0314	0,78

Таблица 2 Сравнение моделей при плотности $\rho_{\text{det}} > 0{,}2$

Модель	Среднеквадратическая ошибка
(5)	**0,058**
(6)	0,119
(7)	0,065
(8)	0,093

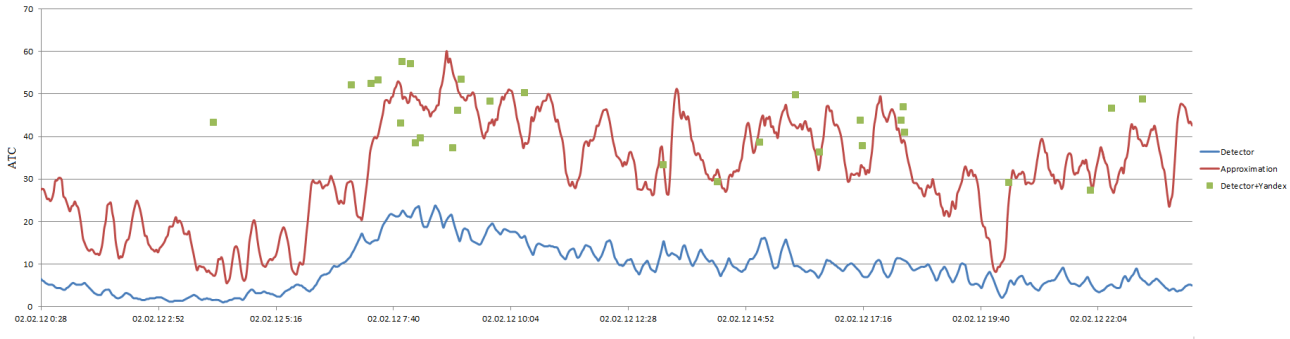


Рис. 7 Пример восстановления числа въехавших АТС. Синяя линия — число проехавших под датчиком на въезде АТС, зеленые точки — сумма данных с датчика и GPS-треков в моменты времени из множества I_{in}. Красная линия — восстановленные значения суммарного числа въехавших АТС N_{estin}

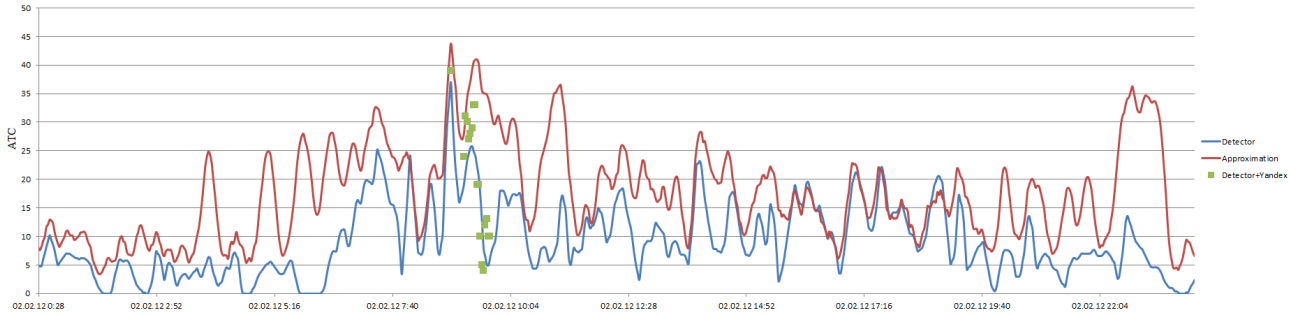


Рис. 8 Пример восстановления числа въехавших АТС. Синяя линия — число проехавших под датчиком на въезде АТС, зеленые точки — сумма данных с датчика и GPS-треков в моменты времени из множества I_{in}. Красная линия — восстановленные значения суммарного числа въехавших АТС N_{estin}

1. Предложен алгоритм восстановления характеристик потока АТС по данным GPS-треков с использованием данных дорожных датчиков на автомагистрали.
2. Проведено сравнение различных моделей оценки числа АТС на автомагистрали с использованием реальных данных и выбрана лучшая.
3. Предложен алгоритм восстановления характеристик потока АТС по данным GPS-треков и дорожных датчиков с учетом возможных значительных временны́х пробелов в них на въездах и съездах с автомагистрали.
4. Проверена работоспособность предложенного алгоритма для восстановления суммарного потока на въездах и съездах на реальных данных.

Особенностью поставленной задачи оценки числа АТС на автомагистрали является большой объем данных с GPS-треков, что позволяет получить приемлемое качество на обучающей и контрольной выборке как для исходной задачи, так и для задачи в реальном времени.

В то же время из-за недостаточного количества данных на въездах и съездах нельзя решить задачу, аналогичную задаче на автомагистрали. Однако, используя уравнение баланса, можно оценить N_{in} и N_{out} на въездах и съездах. В работе предложен алгоритм для получения этих оценок, который дает интерпретируемые результаты.

7 Благодарности

Авторы благодарят компанию Яндекс (ООО «Яндекс») за предоставленные трековые данные, Центр организации дорожного движения за данные датчиков движения. Авторы также чрезвычайно признательны Я. А. Холодову и А. Е. Алексеенко за консультации и обсуждение результатов.

Литература

[1] *Алексеенко А.Е., Холодов Я.А., Холодов А.С., Горева А.И., Васильев М.О., Чехович Ю.В., Мишин В.Д., Старожилец В.М.* Разработка, калибровка и верификация модели движения трафика в городских условиях. Часть I // Компьютерные исследования и моделирование, 2015. Т. 7. №6. P. 1185–1203.

[2] *Бродский Г.С., Кашкин М.Ю., Айвазов А.Р., Рыкунов Р.Р.* Работа детекторов транспорта на Московской дорожно-уличной сети URL: http://www.againc.net/media/3649/dt_article001.pdf.

[3] *Lighthill M.J., Whitham G.B.* On kinematic waves. II. A theory of traffic flow on long crowded roads // P. Roy. Soc. Lond. A Mat., 1955. P. 281–345. doi: 10.1098/rspa.1955.0089.

[4] *Richards P.I.*, Shock waves on the highway // Oper. Res., 1956. Vol. 4. No. 1. P. 42–51. doi: 10.1287/opre.4.1.42.

[5] *Whitham J.B.* Linear and nonlinear waves. — USA: Wiley, 1974. 656 p.

[6] *Daganzo C. F.* Requiem for second-order fluid approximations of traffic flow // Transport. Res. B Meth., 1995. Vol. 29. No. 4. P. 277–286. doi: 10.1016/0191-2615(95)00007-Z.

[7] *Payne H. J.* Models of freeway traffic and control // Math. Models Public Syst., 1998. No. 4. P. 51–61.

[8] *Papageorgiou M.* Some remarks on macroscopic traffic flow modelling // Transport. Res. A Pol., 1998. Vol. 32. No. 5. P. 323–329. doi: 10.1016/S0965-8564(97)00048-7.

[9] *Aw A., Rascle M.* Resurrection of "second order" models of traffic flow // SIAM J. Appl. Math., 2000. Vol. 60. No. 3. P. 916–938. doi: 10.1137/S0036139997332099.

[10] *Zhang M.* A non-equilibrium traffic model devoid of gas-like behavior // Transport. Res. B Meth., 2002. Vol. 36. No. 3. P. 275–290. doi: 10.1016/S0191-2615(00)00050-3.

[11] *Zhang M.* Anisotropic property revisited —does it hold in multi-lane traffic? // Transport. Res. B Meth., 2003. Vol. 37. No. 6. P. 561–577. doi: 10.1016/S0191-2615(02)00030-9.

[12] *Siebel F., Mauser W.* On the fundamental diagram of traffic flow // SIAM J. Appl. Math., 2006. Vol. 66. No. 4. P. 1150–1162. doi: 10.1137/050627113.

[13] *Siebel F., Mauser W.* Synchronized flow and wide moving jams from balanced vehicular traffic // Phys. Rev. E, 2006. Vol. 73. No. 6. P. 066108. doi: 10.1103/PhysRevE.73.066108.

[14] *Воронцов К.В., Чеховић Ю.В.* Интеллектуальный анализ данных в задачах моделирования транспортных потоков // Введение в математическое моделирование транспортных потоков / Под ред. А. В. Гасникова. — М.: Изд-во МЦНМО, 2013. С. 225–248.

[15] *Hall D.L., McMullen S.* Mathematical techniques in multisensor data fusion. — L.: Artech House, 2004. 453 p.

[16] *Воронин А., Михеев Ю.* Синергетические методы комплексирования в задачах принятия решений // 12th Conference (International) "Knowledge–Dialogue–Solution" Proceedings. — Sofia: FOI-COMMERCE, 2006. P. 180–186.

[17] *Khaleghi B., Khamis A., Karray F., Razavi S.* Multisensor data fusion: A review of the state-of-the-art // Inform. Fusion, 2013. Vol. 14. No. 6. P. 28–44. doi: 10.1016/j.inffus.2011.08.001.

[18] *Goodman I., Mahler R., Nguyen H.* Mathematics of data fusion. — New York, NY, USA: Springer Science & Business Media, 2013. 507 p.

[19] *Зенченко В.А., Ременцов А.Н., Павлов А.В., Сотсков А.В.* Оценка параметров окружающей среды и основных транспортных потоков, определяющих ситуацию на улично-дорожной сети // Современные наукоемкие технологии, 2012. № 2. С. 52–59.

[20] *Jerry L., Ray D.* Violin plots: A box plot-density trace synergism // Am. Stat., 1998. Vol. 52. No. 2. P. 181–184. doi: 10.1080/00031305.1998.10480559.

[21] *Kerner B.* The physics of traffic. — Berlin: Springer, 2004. 681 p. doi: 10.1007/978-3-540-40986-1.

Поступила в редакцию 04.08.2016

Aggregation of data from different sources in traffic flow tasks*

***V. M. Starozhilets*[1,2] *and Yu. V. Chekhovich*[1,2]**

starvsevol@gmail.com; chehovich@forecsys.ru

[1]Federal Research Center "Computer Science and Control" of RAS

44/2 Vavilova Str., Moscow, Russia

[2]Moscow Institute of Physics and Technology

9 Institutskiy Per., Dolgoprudny, Moscow Region, Russia

Data aggregation problem, where data are taken from GPS-tracks and traffic detectors, has been studied. Aggregated data are used to state and solve finite differences equation corresponding to the chosen traffic flow mathematical model. The problem is divided into two ones: the first one is about highway data and the second one is about entrances and exits data. To estimate speed and number of cars, a linear model that uses highway data taken from GPS-tracks and traffic detectors has been proposed. The quality criteria are mean squared error and correlation coefficient. Note that the built model can be used on highway data, which do not have data from traffic detectors, but have only data from GPS-tracks. For entrances and

*This research is funded by the Russian Foundation for Basic Research, grant 14-07-00685

exits data, a method to recover summary total flow was developed. This method is based on the preservation of cars in transport network. Computational experiment for both problems is provided on real data and performance of the proposed approaches is demonstrated. Data from GPS-tracks were provided by Yandex.Traffic and data from traffic detectors were provided by Moscow traffic management center. Moscow Ring Road was used as a highway.

Keywords: *data recovery; Moscow Ring Road; GPS-tracks; road detectors*

DOI: 10.21469/22233792.2.3.01

References

[1] Alekseenko, A. E., Ya. A. Kholodov, A. S. Kholodov, A. I. Goreva, M. O. Vasil'ev, Yu. V. Chekhovich, V. D. Mishin, and V. M. Starozhilets. 2015. Razrabotka, kalibrovka verifikatsiya modeli dvizheniya trafika v gorodskikh usloviyakh. Chast' I [Development, calibration and verification of the model of traffic in urban environments. Part I]. *Komp'yuternye issledovaniya i modelirovanie* [Computer Investigations and Simulation] 7(6):1185–1203.

[2] Brodskiy, G. S., M. Yu. Kashkin, A. R. Ayvazov, and R. R. Rykunov. Rabota detektorov transporta na Moskovskoy dorozhno-ulichnoy seti [Traffic detectors working on the Moscow road network]. Available at: http://www.againc.net/media/3649/dt_article001.pdf (accessed November 3, 2016).

[3] Lighthill, M. J., and G. B. Whitham. 1955. On kinematic waves. II. A theory of traffic flow on long crowded roads // *P. Roy. Soc. Lond. A Mat.* 281–345. doi: 10.1098/rspa.1955.0089.

[4] Richards, P. I. 1956. Shock waves on the highway. *Oper. Res.* 4(1):42–51. doi: 10.1287/opre.4.1.42.

[5] Whitham, J. B. 1974. *Linear and nonlinear waves*. USA: Wiley. 656 p.

[6] Daganzo, C. F. 1955. Requiem for second-order fluid approximations of traffic flow. *Transport. Res. B Meth.* 29(4):277–286. doi: 10.1016/0191-2615(95)00007-Z.

[7] Payne, H. J. 1998. Models of freeway traffic and control. *Math. Models Public Syst.* 4:51–61.

[8] Papageorgiou, M. 1988. Some remarks on macroscopic traffic flow modelling. *Transport. Res. A Pol.* 32(5):323–329. doi: 10.1016/S0965-8564(97)00048-7.

[9] Aw, A., and M. Rascle. 2000. Resurrection of "second order" models of traffic flow. *SIAM J. Appl. Math.* 60(3):916–938. doi: 10.1137/S0036139997332099.

[10] Zhang, M.. 2002. A non-equilibrium traffic model devoid of gas-like behavior. *Transport. Res. B Meth.* 36(3):275–290. doi: 10.1016/S0191-2615(00)00050-3.

[11] Zhang, M.. 2003. Anisotropic property revisited — does it hold in multi-lane traffic? // *Transport. Res. B Meth.* 37(6):561–577. doi: 10.1016/S0191-2615(02)00030-9.

[12] Siebel, F., and W. Mauser. 2006. On the fundamental diagram of traffic flow. *SIAM J. Appl. Math.* 66(4):1150–1162. doi: 10.1137/050627113.

[13] Siebel, F., and W. Mauser. 2006. Synchronized flow and wide moving jams from balanced vehicular traffic. *Phys. Rev. E* 73(6):066108. doi: 10.1103/PhysRevE.73.066108.

[14] Vorontsov, K. V., and Yu. V. Chekhovich. 2015. Intellektual'nyy analiz dannykh v zadachakh modelirovaniya transportnykh potokov [Data mining for traffic flows modeling]. *Vvedenie v matematicheskoe modelirovanie transportnykh potokov* [Introduction to the mathematical modeling of traffic flows]. Ed. A. B. Gasnikov. Moscow: MTsNMO. 225–248.

[15] Hall, D. L., and S. McMullen. 2004. *Mathematical techniques in multisensor data fusion*. L.: Artech House. 453 p.

[16] Voronin, A., and Yu. Mikheev. 2006. Sinergeticheskie metody kompleksirovaniya v zadachakh prinyatiya resheniy [Synergetic methods of aggregation in decision problems]. *12th Conference (International) "Knowledge–Dialogue–Solution" Proceedings*. Sofia: FOI-COMMERCE. P. 180–186.

[17] Khaleghi, B., A. Khamis, F. Karray, and S. Razavi. 2013. Multisensor data fusion: A review of the state-of-the-art. *Inform. Fusion* 14(6):28–44. doi: 10.1016/j.inffus.2011.08.001.

[18] Goodman, I., R. Mahler, and H. Nguyen. 2013. *Mathematics of data fusion*. New York, NY: Springer Science & Business Media. 507 p.

[19] Zenchenko, V. A., A. N. Rementsov, A. V. Pavlov, and A. V. Sotskov. 2012. Otsenka parametrov okruzhayushchey sredy i osnovnykh transportnykh potokov, opredelyayushchikh situatsiyu na ulichno-dorozhnoy seti [Evaluation of the environmental parameters and the main traffic flows that determine the situation on the road network]. *Sovremennye Naukoemkie Tekhnologii* [Modern High Technologies] 2:52–59.

[20] Jerry, L., and D. Ray. 1998. Violin plots: A box plot-density trace synergism. *Am. Stat.* 52(2):181–184. doi: 10.1080/00031305.1998.10480559.

[21] Kerner, B. 2004. *The physics of traffic*. Berlin: Springer. 681 p. doi: 10.1007/978-3-540-40986-1.

Received August 4, 2016

Решающие правила для ансамбля из цепей вероятностных классификаторов при решении задач классификации с пересекающимися классами

А. А. Остапец
aostapec@mail.ru
МГУ им. М. В. Ломоносова, Россия, г. Москва, Ленинские горы, д. 1

Рассматривается задача классификации с пересекающимися классами. Исследовано применение ансамбля из цепей вероятностных классификаторов с использованием основных типов решающих правил для формирования итоговых предсказаний. Схема решения рассматривается с точки зрения алгебраического подхода. Алгебраический подход заключается в представлении алгоритма решения задачи в виде суперпозиции двух алгоритмов. На первом этапе строится первый алгоритм (распознающий оператор), который в качестве ответа выдает вектор оценок принадлежности к каждому из классов. В качестве распознающих операторов рассматриваются следующие семейства алгоритмов: линейные классификаторы (базовые классификаторы), цепь вероятностных классификаторов из линейных классификаторов и ансамбль из цепей вероятностных классификаторов. На следующем этапе второй алгоритм (решающее правило) трансформирует этот вектор оценок в финальный ответ. Приведен обзор основных типов решающих правил и исследовано их применение для различных распознающих операторов. Экспериментально показана возможность эффективного использования решающих правил, построенных над результатами прогнозов базовых классификаторов.

Ключевые слова: *решающие правила; классификация с пересекающимися классами; построение ансамблей; классификация текстов*

DOI: 10.21469/22233792.2.3.02

1 Введение

Задача классификации с непересекающимися классами является широко распространенной среди задач машинного обучения. В этой задаче каждый объект связан ровно с одним целевым классом. В зависимости от числа непересекающихся классов $\mathcal{L}$ различают задачу бинарной классификации (при $|\mathcal{L}| = 2$) и задачу многоклассовой классификации (при $|\mathcal{L}| > 2$). Задача классификации с пересекающимися классами позволяет объектам относиться к нескольким классам одновременно.

Пусть $\mathcal{X}$ — пространство объектов; $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ — конечное множество классов; $\mathcal{Y} = \{0, 1\}^k$ — множество всех бинарных векторов размерности k. Объект $\mathbf{x} \in \mathcal{X}$ описывается вектором признаков $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ и принадлежит некоторому подмножеству классов L из $\mathcal{L}$. Сопоставим каждому подмножеству классов L бинарный вектор $\mathbf{Y} = (y_1, y_2, \ldots, y_k)$, где $y_i = 1 \Leftrightarrow \lambda_i \in L$.

Определение 1. *Пусть дан тренировочный набор данных $S = (\mathbf{x}_i, \mathbf{Y}_i), 1 \leqslant i \leqslant n$, состоящий из n объектов ($\mathbf{x}_i \in \mathcal{X}, \mathbf{Y}_i \in \mathcal{Y}$), взятых из неизвестного распределения D. Алгоритмом для решения задачи с пересекающимися классами является классификатор $h : \mathcal{X} \to \mathcal{Y}$, который оптимизирует заданную функцию потерь* [1].

Часто алгоритмы для решения задачи с пересекающимися классами вместо бинарной классификации для каждого класса, определенной выше, представляют собой вещественнозначную функцию $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$. С помощью данной функции для классифицируемого объекта формируется вектор оценок принадлежности $(g_1, \ldots, g_k)$ к классам $\mathcal{L}$. Каждый

такой алгоритм нуждается во втором алгоритме (решающем правиле), который трансформирует этот вектор оценок $(g_1, \ldots, g_k)$ в бинарный вектор $(a_1, \ldots, a_k) \in \mathcal{Y}$. Ненулевые элементы этого вектора — это множество классов, к которым алгоритм относит объект.

2 Методы преобразования задачи

Существуют несколько достаточно простых методов преобразования, которые переводят задачу классификации с пересекающимися классами в задачу, к которой могут быть применены существующие алгоритмы многоклассовой классификации. В данной работе, в дальнейшем, рассматривается метод Binary Relevance (BR), а методы Label Powerset (LP) и Error-Correcting Output Code (ECOC) приведены в качестве альтернативных вариантов решения проблемы преобразования задачи.

2.1 Label Powerset

Label Powerset — это простой метод, который рассматривает каждое уникальное множество классов в исходной обучающей выборке как один новый класс в преобразованных данных. К преобразованной задаче могут применяться любые алгоритмы многоклассовой классификации. Предсказанный алгоритмом класс в новой задаче однозначно соответствует определенному множеству классов в исходной задаче. Благодаря методу LP также можно осуществлять ранжирование по вероятности исходных классов при предсказании, используя оценки классификатора на новых сформированных классах [2].

Одна из проблем метода LP заключается в том, что после преобразования данных большая часть новых классов содержит очень мало объектов и распределение объектов в новых классах является крайне несбалансированным. Для решения этой проблемы был предложен метод [2]. В этом методе первым делом подбирается порог для отсечения и находятся все классы, которые в преобразованных данных имеют частоту ниже этого порога. Каждый из таких классов заменяется на меньшие по мощности, непересекающиеся подмножества из исходных классов. Каждое из подмножеств должно иметь частоту выше установленного порога в преобразованных данных.

2.2 Binary Relevance

Binary Relevance [3] — это один из самых популярных методов преобразования задачи классификации с пересекающимися классами в задачу с непересекающимися классами. Этот метод создает k наборов данных ($k = |\mathcal{L}|$), по одному набору данных на каждый класс. Все новые наборы данных содержат одинаковое число объектов, равное числу объектов в исходной обучающей выборке. В каждом наборе данных $D_{\lambda_j}, 1 \leqslant j \leqslant k$, позитивным классом являются объекты, которые принадлежат классу λ_j, а негативный класс присваивается всем оставшимся объектам.

На каждом наборе данных обучается бинарный классификатор. На этапе предсказания для объекта берутся предсказания от каждого бинарного классификатора. Итоговым ответом является объединение классов λ_j, которые бинарные классификаторы определили как позитивные для объекта. Несмотря на то что BR подход используется во многих практических приложениях, он часто критикуется за неявное предположение о независимости исходных классов, которое может не выполняться на реальных данных.

2.3 Error-Correcting Output Code

Интересный метод преобразования задачи многоклассовой классификации в несколько задач бинарной классификации был предложен в работе [4]. Этот метод получил название Error-Correcting Output Code. Алгоритм преобразования заключается в кодировании

меток классов бинарными векторами длины l. После этого задача нахождения истинного класса для объекта $\mathbf{x} \in \mathcal{X}$ сводится к определению l неизвестных бит кодового слова класса $y(x)$. Для каждого бита i строится бинарный классификатор f_i, отделяющий группу классов со значением $+1$ соответствующего бита от классов со значением -1. При классификации для объекта $\mathbf{x}$ вычисляется кодовое слово $\mathbf{f}(x) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_l(\mathbf{x}))$ и выбирается ближайший к $\mathbf{f}(x)$ по расстоянию Хэмминга класс. Для задач с пересекающимися классами этот подход можно совместить с подходом LP.

2.4 Использование связей между классами

Многие методы преобразования (например, рассмотренный выше BR имеют в себе предположение о независимости классов. Одна из идей для оценки совместного распределения классов была предложена в [5]. Для этого метода необходимо обучить k различных функций ($k = |\mathcal{L}|$) на расширенных признаковых пространствах $\mathcal{X} \times \{0,1\}^{i-1}$, где $y_1, y_2, \ldots, y_{i-1}$ — это оценки принадлежности к классам $\lambda_1, \lambda_2, \ldots, \lambda_{i-1}$;

$$f_i : \mathcal{X} \times \{0,1\}^{i-1} \to [0,1]\,;$$
$$(\mathbf{x}, y_1, y_2, \ldots, y_{i-1}) \to \mathsf{P}(y_i = 1 | \mathbf{x}, y_1, y_2, \ldots, y_{i-1})\,.$$

Здесь f_i могут интерпретироваться как вероятностные классификаторы. Этот подход называется цепью вероятностных классификаторов (Probabilistic Classifier Chain — PCC). В работе [5] описано, как создать ансамбль из таких классификаторов (Ensembled PCC). На каждой итерации новый построенный алгоритм будет отличаться от остальных:

- выбором случайного подмножества объектов для обучения;
- случайной перестановкой порядка классов $\lambda_1, \lambda_2, \ldots, \lambda_k$.

3 Решающие правила

Алгебраический подход [6, 7] заключается в представлении алгоритма решения задачи в виде суперпозиции двух алгоритмов:

1) первый алгоритм (распознающий оператор) строит вектор оценок принадлежности классам $(g_1, \ldots, g_k)$, где g_j — оценка принадлежности объекта к j-му классу;
2) второй алгоритм (решающее правило) трансформирует вектор оценок $(g_1, \ldots, g_k)$ в бинарный вектор $(a_1, \ldots, a_k) \in \{0,1\}^k$. Ненулевые элементы этого вектора — это классы, к которым алгоритм относит объект.

В работе [8] представлены 4 вида решающих правил. Описание указанных решающих правил приводится ниже.

3.1 S-cut

Простейшее решающее правило, возвращающее множество классов, которые получили оценку принадлежности не ниже, чем заданный константный порог t:

$$a_i(\mathbf{x}) = \mathbb{I}[g_i(\mathbf{x}) \geqslant t], \forall i \in \mathcal{L}\,.$$

Оптимальное значение порога t можно определить, например, с помощью кросс-валидации.

3.2 R-cut

Решающее правило R-cut, в отличие от предыдущего решающего правила, всегда возвращает в качестве ответа множество ровно из r классов с наивысшими оценками:

$$a_i(\mathbf{x}) = \mathbb{I}[\mathrm{rank}(i) \leqslant r], \forall i \in \mathcal{L},$$

где rank(i) — это соответствующая $g_i(\mathbf{x})$ позиция класса i в отсортированным по невозрастанию списке оценок. Как и в случае S-cut, оптимальное значение параметра r можно выбрать с помощью кросс-валидации. Интересный вариант этой стратегии представлен в работе [9], где вместо константного значения r количество классов определяется индивидуально для каждого объекта.

3.3 DS-cut

В этом решающем правиле используются несколько параметров t_i, каждый из них соответствует позиции i, которую класс занимает в отсортированном списке оценок:

$$a_i(\mathbf{x}) = \mathbb{I}[g_i(\mathbf{x}) \geqslant t_{\text{rank}(i)}], \forall i \in \mathcal{L}\,.$$

Здесь необходимо определять не один параметр, как в предыдущих решающих правилах, а p параметров (можно считать, что $t_{(p+1)} = \cdots = t_k = +\infty$). Таким образом, алгоритм может вернуть не более p классов (обычно значение p выбирается равным от 4 до 7).

3.4 DSS-cut

Это решающее правило похоже на предыдущее, за исключением того, что абсолютные значения оценок заменены на отношение оценок на конкретных позициях и максимальной оценки в векторе. Эти отношения сравниваются с порогами:

$$a_i(\mathbf{x}) = \mathbb{I}[\frac{g_i(\mathbf{x})}{g_{\max}} \geqslant t_{\text{rank}(i)}], \forall i \in \mathcal{L}\,,$$

где g_{max} — это максимальная оценка в векторе $(g_1, \ldots, g_k)$. В случае если t_1 — порог для максимальной оценки — установлен равным 1, то алгоритм всегда возвращает по меньшей мере один класс для каждого объекта. Аналогично предыдущему правилу алгоритм может вернуть не более p классов.

Отметим, что для всех рассмотренных выше правил можно подбирать не глобальные пороги, а пороги для каждого класса в отдельности.

4 Эксперименты

4.1 Набор данных

Эксперименты проводились на наборе данных, предложенном участникам конкурса Greek Media Monitoring Multilabel Classification (WISE 2014) [10], который проводился на платформе Kaggle летом 2014 г. Данные представляли собой статьи, которые были размещены в греческих средствах массовой информации в период с мая по сентябрь 2013 г. Текст каждой статьи был представлен с использованием модели «мешок слов», после чего было осуществлено TF–IDF (term frequency – inverse document frequency) преобразование. Суть модели «мешок слов» состоит в том, что в ней учитывается только количество вхождений каждого слова в документ, а порядок слов в документе полностью игнорируется. TF-IDF — это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес слова в этом преобразовании пропорционален количеству употреблений этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции.

Таким образом, в исходных данных не предоставлены оригинальные тексты, а только предпросчитанные признаки для каждой статьи. Эксперты в данной задаче вручную

разметили классы для этого набора данных. Каждая статья может принадлежать одному или нескольким классам из 203 доступных.

4.2 Функционалы качества

В качестве функционалов качества в данной статье будут рассматриваться 2 варианта:

1) усредненная по всем объектам F-мера;
2) точность классификации.

Усредненная по всем объектам F-мера (Mean F1-Score, Example-Based F-measure) часто применяется в задачах с пересекающимися классами. При вычислении этого функционала для каждого объекта вычисляется значение F-меры, а затем все полученные значения усредняются:

$$F_{\text{score}} = \frac{1}{M}\sum_{i=1}^{M} f^i_{\text{score}}, \quad f^i_{\text{score}} = 2\frac{pr}{p+r}$$

где

$$p = \frac{\text{tp}}{\text{tp}+\text{fp}}; \quad r = \frac{\text{tp}}{\text{tp}+\text{fn}};$$

tp — количество истинно-положительных значений для объекта i; fp — количество ложно-положительных значений для объекта i; fn — количество ложно-отрицательных значений для объекта i; M — число объектов в тестовой выборке.

При вычислении данной метрики объект i можно представить множеством классов, к которым этот объект действительно принадлежит. Вторым множеством будет множество классов, которые были предсказаны для этого объекта. Тогда для объекта i характеристики tp, fp и fn вычисляются следующим образом:

1) tp — пересечение двух множеств, описанных выше;
2) fp — предсказанные классы, к которым объект не принадлежит в действительности;
3) fn — классы, к которым объект принадлежит в действительности, но которые отсутствуют в множестве предсказанных классов.

В задачах с пересекающимися классами под **точностью классификации** обычно понимают полное совпадение множеств (subset accuracy): предсказанное множество классов должно полностью совпадать с множеством истинных классов.

Пусть Y_i^{pred} — бинарный вектор принадлежности i-го объекта k классам, полученный алгоритмом, а Y_i^{true} — бинарный вектор принадлежности i-го объекта k классам из тестовой выборки. Тогда данный функционал будет вычисляться следующим образом:

$$\text{Accuracy} = \frac{1}{M}\sum_{i=1}^{M} \text{acc}\left(Y_i^{\text{pred}}, Y_i^{\text{true}}\right),$$

где

$$\text{acc}\left(Y_i^{\text{pred}}, Y_i^{\text{true}}\right) = \begin{cases} 1, & \text{если } Y_i^{\text{pred}} \text{ полностью совпадает с } Y_i^{\text{true}}; \\ 0 & \text{иначе.} \end{cases}$$

4.3 Результаты

Для экспериментов были зафиксированы следующие модели:

– логистическая регрессия (ЛР) (использовалась реализация из scikit-learn [11] с параметрами penalty='l1', C=6,0 и tol=0,001;

Таблица 1 Mean F1-Score для различных решающих правил

Алгоритм	S-cut	R-cut	DS-cut	DSS-cut
Логистическая регрессия	73,07	73,58	76,36	78,28
Одна модель PCC на основе ЛР	73,99	73,40	76,27	78,24
Две модели PCC на основе ЛР	74,52	73,68	76,68	78,32
Три модели PCC на основе ЛР	74,48	73,73	76,74	78,41
Линейный классификатор (SGD)	71,80	71,53	71,12	75,52
Одна модель PCC на основе ЛК	71,96	71,46	71,06	75,41
Две модели PCC на основе ЛК	72,13	71,66	71,41	75,55
Три модели PCC на основе ЛК	72,18	71,78	71,50	75,67

– линейный классификатор, обученный с помощью метода стохастического градиента (Stochastic Gradient Descent — SGD) (использовалась реализация из scikit-learn [11]: SGDClassifier(loss="modified_huber")).

На основе каждой из этих моделей строились 4 распознающих оператора:

1) исходная модель, обученная с помощью метода BR;
2) цепь вероятностных классификаторов исходной модели;
3) ансамбль из двух цепей вероятностных классификаторов исходной модели;
4) ансамбль из трех цепей вероятностных классификаторов.

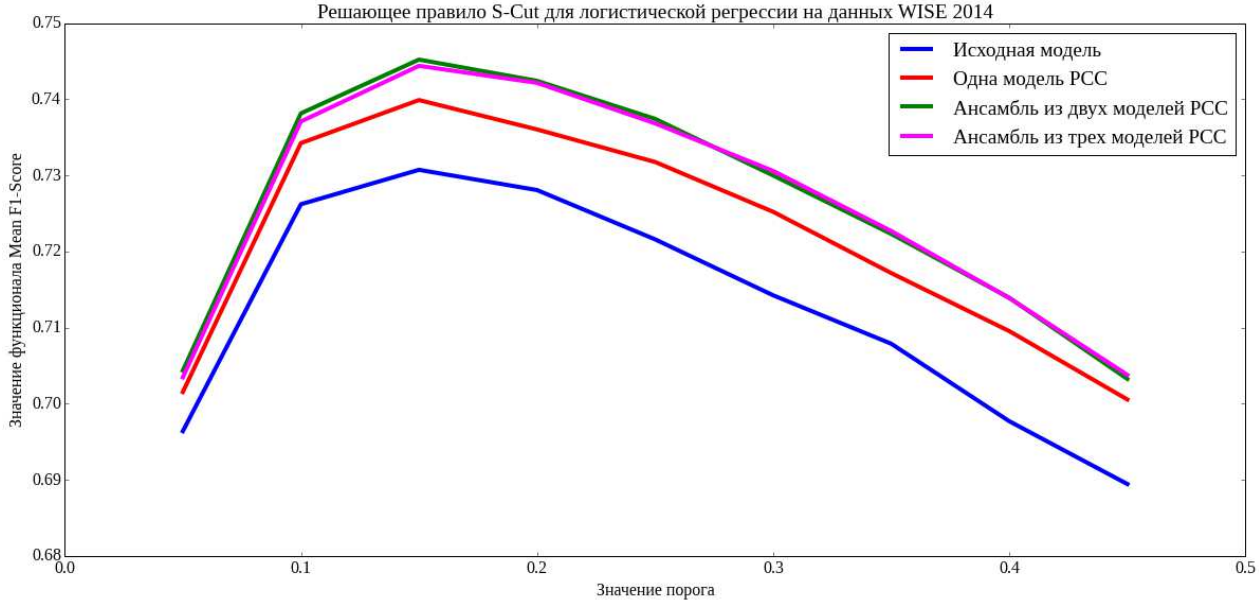
Для всех этих распознающих операторов подбирались пороги для рассмотренных выше решающих правил.

В табл. 1 показаны лучшие значения Mean F1-Score для различных решающих правил (т. е. для каждого решающего правила перебиралось множество различных значений, и лучший результат отображен в таблице). Отметим, что исходную модель для всех решающих правил начинает обходить только ансамбль из двух цепей вероятностных классификаторов.

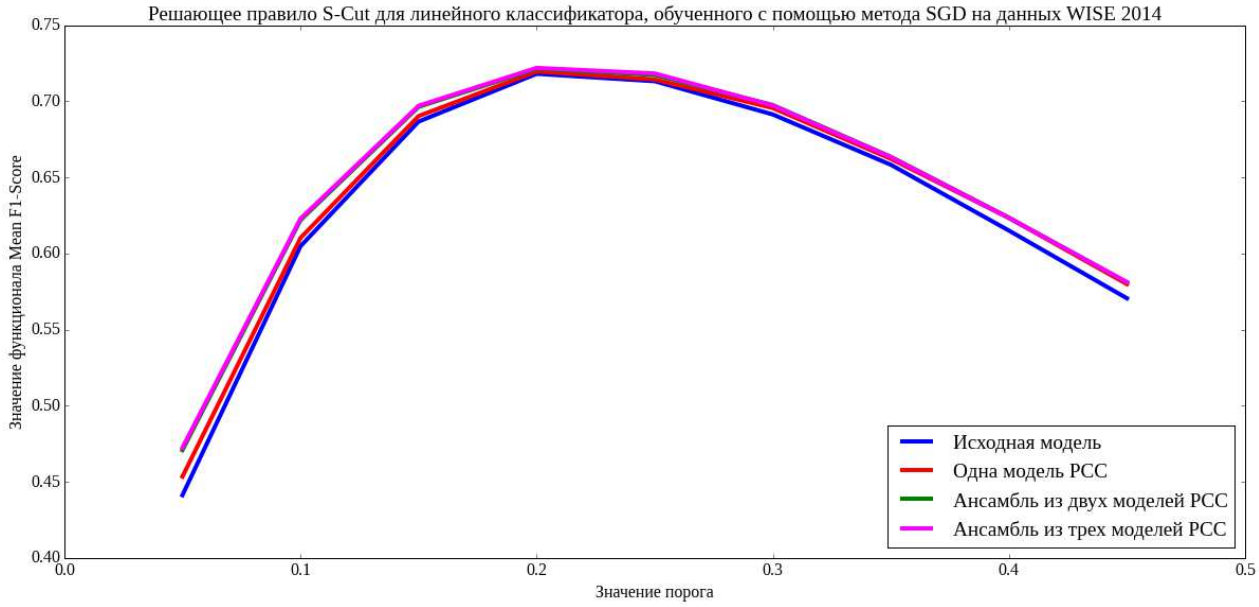
На рис. 1, *а* показана зависимость величины Mean F1-Score решения от порога для решающего правила S-cut для ЛР. Видим, что для этого решающего правила цепь вероятностных классификаторов начинает работать лучше исходной модели даже без использования ансамблей. На рис. 1, *б* показана зависимость качества решения от порога для решающего правила S-cut для линейного классификатора. Здесь качество всех моделей схоже и выигрыш от использования ансамбля значительно меньше.

В табл. 2 показаны лучшие значения точности классификации для различных решающих правил (как и в предыдущем функционале, для каждого решающего правила перебиралось множество различных значений, и лучший результат отображен в таблице).

На рис. 2, *а* показана зависимость величины точности классификации от порога для решающего правила S-cut для ЛР. Как и в предыдущем случае, для этого решающего правила цепь вероятностных классификаторов начинает работать лучше исходной модели даже без использования ансамблей. На рис. 2, *б* показана зависимость величины точности классификации от порога для решающего правила S-cut для линейного классификатора. Аналогично рис. 1, *б* качество всех моделей схоже и выигрыш от использования ансамбля значительно меньше.



(*а*)

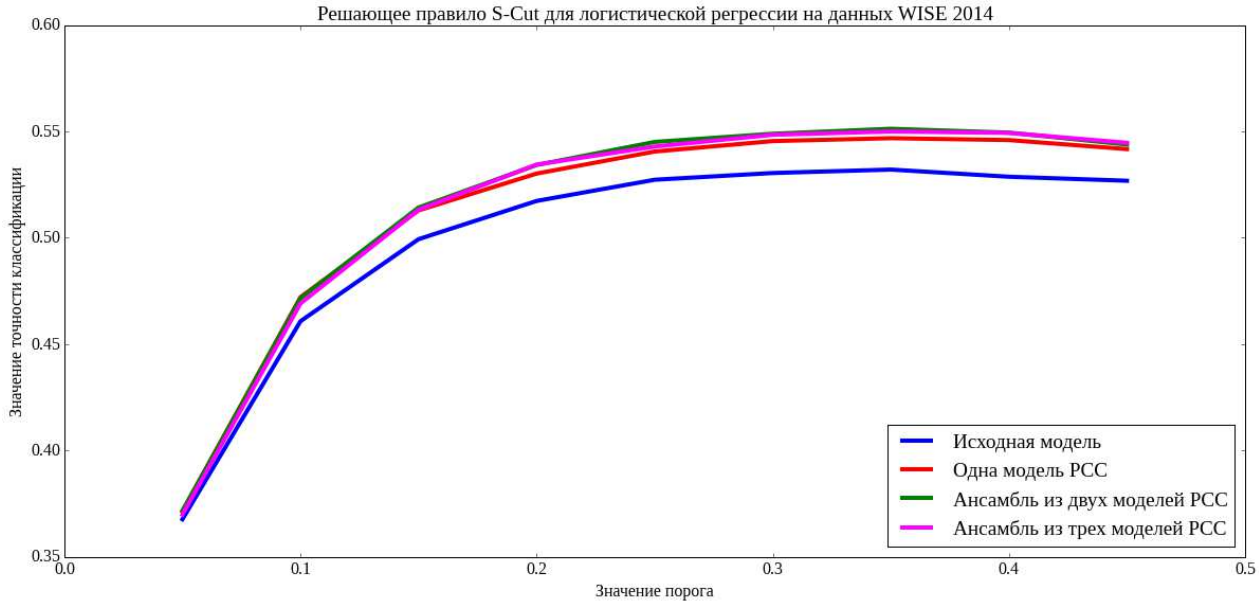


(*б*)

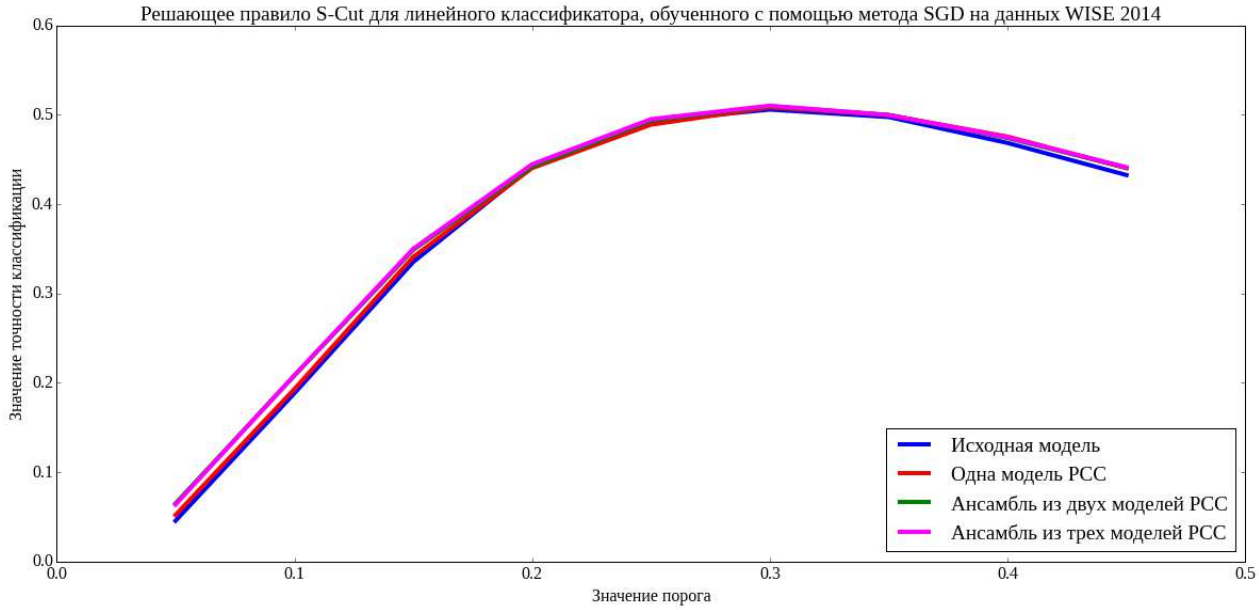
Рис. 1 Решающее правило S-cut для ЛР (*а*) и линейного классификатора (*б*) (Mean F1-Score)

Таблица 2 Точность классификации для различных решающих правил

Алгоритм	S-cut	R-cut	DS-cut	DSS-cut
Логистическая регрессия	52,73	58,29	53,77	59,93
Одна модель PCC на основе ЛР	54,68	58,17	54,00	59,85
Две модели PCC на основе ЛР	55,13	58,42	54,19	60,15
Три модели PCC на основе ЛР	55,20	58,50	54,25	60,21
Линейный классификатор (SGD)	50,58	56,77	53,40	53,20
Одна модель PCC на основе ЛК	50,82	56,62	53,32	53,18
Две модели PCC на основе ЛК	50,94	56,89	53,51	53,55
Три модели PCC на основе ЛК	51,00	56,96	53,64	53,73



(a)



($б$)

Рис. 2 Решающее правило S-cut для ЛР (a) и линейного классификатора ($б$) (точность классификации)

5 Заключение

Показано, что благодаря использованию достаточно простой идеи об учете взаимосвязи между классами и подбора верного решающего правила удается улучшать качество работы базовой модели, обученной с помощью метода BR.

Отдельно отметим, что построение одной цепи вероятностных классификаторов не приводит к улучшению качества работы исходной модели. Необходимый прирост в качестве дает использование ансамбля из двух и более цепей вероятностных классификаторов.

Литература

[1] *Zhang M. L., Zhou Z. H.* ML-KNN: A lazy learning approach to multi-label learning // Pattern Recogn., 2007. Vol. 40. No. 7. P. 2038–2048.

[2] *Read J.* A pruned problem transformation method for multi-label classification // 2008 New Zealand Computer Science Research Student Conference Proceedings. P. 143–150.

[3] *Godbole S., Sarawagi S.* 2004. Discriminative methods for multi-labeled classification // PAKDD '04: 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining. — Springer, 2008. P. 22–30.

[4] *Dietterich T. G., Bakiri G.* Solving multiclass learning problems via error-correcting output codes // J. Artificial Intell. Res., 1995. Vol. 2. P. 263–286.

[5] *Dembczynski K., Cheng W., Hullermeier E.* Bayes optimal multilabel classification via probabilistic classifier chains. 27th Conference (International) on Machine Learning. — Haifa, Israel, 2010. P. 279–286.

[6] *Журавлёв Ю. И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов. III // Кибернетика, 1978. №2. С. 35–43.

[7] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания // Проблемы кибернетики, 1979. Вып. 33. С. 5–68.

[8] *Wang X. L., Zhao H., Lu B. L.* Enhanced K-nearest neighbour algorithm for large-scale hierarchical multi-label classification // Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classification Proceedings, 2011.

[9] *Quevedo J. R., Luaces O., Bahamonde A.* Multilabel classifiers with a probabilistic thresholding strategy // Pattern Recogn., 2012. Vol. 45. No. 2 P. 876–883.

[10] Greek Media Monitoring Multilabel Classification, 2014. `http://www.kaggle.com/c/wise-2014`.

[11] Scikit-learn: Machine learning in Python. `http://scikit-learn.org`.

Поступила в редакцию 28.08.2016

Decision rules for ensembled probabilistic classifier chain for multilabel classification

A. A. Ostapets

aostapec@mail.ru

Lomonosov Moscow State University, 1 Leninskie Gory, Moscow, Russia

This work considers using of the main types of decision rules for the multilabel classification task. The algorithm is presented as a superposition of two algorithms: a recognition operator and a decision rule. The recognition operator converts feature vectors of objects to be recognized into scores for each class. This work considers several families of algorithms to be the recognition operator: linear models (base classifiers), probabilistic classifier chain of linear models, and ensembled probabilistic classifier chain. The decision rule converts the scores into the final answers. In this survey, main types of decision rules are described and their performance for several recognition operators is also shown. It is experimentally demonstrated that the quality of the forecast of the proposed composition exceeds the quality of the base classifiers.

Keywords: *decision rules; multilabel classification; building ensembles; text classification*

DOI: 10.21469/22233792.2.3.02

References

[1] Zhang, M. L., and Z. H. Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recogn.* 40(7):2038–2048.

[2] Read, J. 2008. A pruned problem transformation method for multi-label classification. *2008 New Zealand Computer Science Research Student Conference Proceedings.* 143–150.

[3] Godbole, S., and S. Sarawagi. 2004. Discriminative methods for multi-labeled classification. *PAKDD '04: 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining.* Springer. 22–30.

[4] Dietterich, T. G., and G. Bakiri. 1995. Solving multiclass learning problems via error-correcting output codes. *J. Artificial Intell. Res.* 2:263–286.

[5] Dembczynski, K., W. Cheng, and E. Hullermeier. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. *27th Conference (International) on Machine Learning.* Haifa, Israel. 279–286.

[6] Zhuravlev, Yu. I. 1978. Korrektnye algebry nad mnozhestvom nekorrektnykh (evristicheskikh) algoritmov. III [Correct algebras for sets of incorrect (heuristic) algorithms. III]. *Kibernetika* [Cybernetics] 2:35–43.

[7] Zhuravlev, Yu. I. 1979. Ob algebraicheskom podkhode k resheniyu zadach raspoznavaniya [An algebraic approach to recognition and classification problems]. *Problemy kibernetiki* [Problems of Cybernetics] 33:5–68.

[8] Wang, X. L., H. Zhao, and B. L. Lu. 2011. Enhanced K-nearest neighbour algorithm for large-scale hierarchical multi-label classification. *Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classification Proceedings.*

[9] Quevedo, J. R., O. Luaces, and A. Bahamonde. 2012. Multilabel classifiers with a probabilistic thresholding strategy. *Pattern Recogn.* 45(2):876–883.

[10] Greek Media Monitoring Multilabel Classification. 2014. Available at: `http://www.kaggle.com/c/wise-2014` (accessed December 13, 2016).

[11] Scikit-learn: Machine learning in Python. Available at: `http://scikit-learn.org` (accessed December 13, 2016).

Received August 28, 2016

Метрики на основе оптимального выравнивания биомолекулярных последовательностей*

***В. В. Сулимова*[1], *О. С. Середин*[1], *В. В. Моттль*[2]**

vsulimova@yandex.ru; oseredin@yandex.ru; vmottl@yandex.ru

[1]ФГБОУ ВО Тульский государственный университет, Россия, г. Тула, пр. Ленина, д. 92

[2]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, д. 44/2

Для биомолекулярных последовательностей наиболее адекватным является так называемый беспризнаковый подход, основанный на сравнении последовательностей (измерении их сходства или несходства), минуя явное вычисление векторов их признаков. С точки зрения передовых методов анализа данных наиболее предпочтительным является использование в качестве способа сравнения меры несходства, обладающей свойствами метрики. С другой стороны, с точки зрения молекулярной биологии важно, чтобы способ сравнения учитывал биологические особенности объектов сравнения. Кроме того, в условиях обработки больших объемов данных важно, чтобы способ сравнения был эффективен с вычислительной точки зрения и позволял в дальнейшем применять удобные и эффективные методы анализа данных, такие как метод опорных векторов (SVM — support vector machine). Известно множество способов сравнения биомолекулярных последовательностей, однако ни один из них не обладает всеми требуемыми свойствами. В данной работе предлагается достаточно простой способ построения метрик на множестве биомолекулярных последовательностей. Предлагаемый метод, как и традиционные общепринятые способы сравнения биомолекулярных последовательностей (такие, как алгоритм Нидлмана–Вунша и Смита–Ватермана), основывается на поиске их оптимального парного выравнивания и механизме мутационных замен аминокислот в ходе эволюции, но отличается от них используемым критерием оптимальности, типом оптимизации и способом сравнения элементов последовательностей. Приводится доказательство того, что предложенные меры несходства обладают свойствами метрики. Это позволяет использовать их в передовых методах анализа данных, сохраняющих вычислительные достоинства SVM, но не требующих введения признаков последовательностей и(или) скалярного произведения. Результаты экспериментов подтверждают адекватность предложенных метрик прикладным задачам на примере классификации мембранных гликопротеинов.

Ключевые слова: *метрики; сравнение последовательностей; оптимальное парное выравнивание; биомолекулярные последовательности; беспризнаковый подход*

DOI: 10.21469/22233792.2.3.03

1 Введение

Биомолекулярные последовательности, к которым относят нуклеотидные и аминокислотные последовательности, образующие полимерные молекулы белка, являются типовыми объектами анализа данных. Основной целью их анализа является определение заключающейся в них генетической информации и функций, которые они выполняют в организме. Результаты анализа биомолекулярных последовательностей крайне важны и находят применение в медицине, фармакологии, косметологии, биотехнологии, сельском

*Работа выполнена при финансовой поддержке РФФИ, проект №15-07-08967.

хозяйстве, экологии и других областях. В частности, они используются при изучении молекулярных механизмов болезней, выявлении предрасположенности человека к заболеваниям, разработке новых лекарственных средств и т. д.

Для анализа биомолекулярных последовательностей необходимо уметь сравнивать их между собой.

Традиционными способами сравнения биомолекулярных последовательностей являются меры сходства, основанные на оптимальном парном выравнивании [1–4]. Однако такие способы сравнения не позволяют при дальнейшем анализе использовать преимущества удобных и эффективных линейных методов анализа данных, разработанных для признаковых пространств, например хорошо зарекомендовавшего себя SVM [5].

В ряде случаев для обеспечения возможности применения SVM осуществляется искусственное введение так называемых вторичных (проекционных) признаков [6–11]. Однако при этом происходит искажение исходного биологически обоснованного понимания сходства последовательностей, что с точки зрения молекулярной биологии является нежелательным. Кроме того, такой подход требует знания и запоминания значений парного сходства для всех исследуемых последовательностей, что вносит существенные неудобства с вычислительной точки зрения, нейтрализуя вычислительные достоинства SVM.

Отчасти эту проблему решает использование специальной меры сходства, называемой потенциальной функцией (kernel function) [9, 12, 13]. Потенциальная функция, определенная на множестве объектов произвольной природы, погружает это множество в гипотетическое линейное пространство, в котором играет роль скалярного произведения [14]. Построению таких функций посвящено множество публикаций (см., например, [9, 13, 15–20]). Однако вычисление потенциальных функций, являющихся не только математически корректными, но и имеющими смысл с точки зрения молекулярной биологии, является непростой и, как правило, вычислительно очень трудоемкой задачей [13, 18, 21]).

В то же время, несложно убедиться, что понятие линейного пространства является избыточным. Все известные методы анализа данных опираются именно на метрику (т. е. взаимное расположение объектов в пространстве) и именно метрика, а не координаты объектов в линейном пространстве определяют в конечном счете результат решения задачи [22, 23], что хорошо видно на рис. 1. Более того, существуют целые классы потенциальных функций, порождающих одну и ту же метрику и, соответственно, являющихся эквивалентными с точки зрения получаемых решений [22, 23].

В связи с этим гораздо более естественным представляется в качестве способа сравнения использовать меру несходства, обладающую свойствами метрики, тем более что последние исследования в области беспризнакового анализа данных показывают, что в терминах метрических пространств могут быть сформулированы многие методы анализа данных, включая SVM с сохранением его основных достоинств и свойств [24–28].

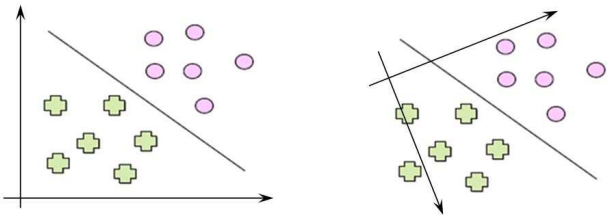


Рис. 1 Два класса объектов и решающее правило их распознавания в линейных пространствах, отличающихся выбором центра и базиса

При этом очевидно, что не любые метрики, построенные на множестве последовательностей, позволят получить приемлемое качество решения задачи анализа последовательностей, а только те, которые удовлетворяют так называемой гипотезе компактности — последовательности, относящиеся к одному классу (например, выполняющие одинаковую функцию) отображаются в более близкие точки данного пространства по сравнению с последовательностями, выполняющими разные функции.

В литературе известен ряд способов введения метрики на множестве последовательностей [29]. Однако они не имеют интерпретации с точки зрения молекулярной биологии, в связи с чем выполнение гипотезы компактности в порождаемом ими пространстве представляется маловероятным. Это находит многократные подтверждения на практике [30, 31] и порождает целую серию публикаций, направленных на поиск способов улучшения исходной метрики (либо ее формирования на основе меры сходства) при помощи алгебраических конструкций различной степени сложности (Metric Learning) [30–34], включая построение метрик на основе потенциальных функций (Metric Kernel Learning) и проекционных признаков, что имеет описанные выше недостатки.

В данной работе предлагается достаточно простой способ сравнения биомолекулярных последовательностей, основанный, как и традиционные методы, на поиске оптимального парного выравнивания сравниваемых последовательностей, однако используются другой критерий оптимальности и другой способ сравнения элементов. В работе приводится доказательство, что предлагаемая функция парного сравнения обладает свойствами метрики. Экспериментальное исследование показывает, что данная метрика может успешно применяться для анализа биомолекулярных последовательностей.

2 Метрики на множестве элементов биомолекулярных последовательностей

Очевидно, что сравнение последовательностей должно базироваться на сравнении составляющих их элементов.

Типичным примером биомолекулярных последовательностей являются аминокислотные последовательности белков, т. е. последовательности над алфавитом двадцати известных аминокислот $A = \{\alpha^1, \ldots, \alpha^m\}, m = 20$.

В качестве теоретической концепции сравнения аминокислот в данной работе принята вероятностная модель эволюции Маргарет Дэйхофф, называемая PAM (Pointed Accepted Mutation) [35]. Ее основным математическим понятием является понятие марковской цепи эволюции аминокислот в отдельно взятой точке цепи, определяемой матрицей переходных вероятностей $\mathbf{\Psi} = \left(\psi_{[1]}(\alpha^j|\alpha^i)\right)$ замены аминокислоты α^i на аминокислоту α^j на следующем шаге эволюции. При этом предполагается, что эта марковская цепь представляет собой эргодический и обратимый случайный процесс, т. е. процесс, характеризующийся финальным распределением вероятностей $\xi(\alpha^j)$:

$$\sum_{\alpha^i \in A} \xi\left(\alpha^i\right)\psi_{[1]}\left(\alpha^j|\alpha^i\right) = \xi\left(\alpha^j\right)$$

и удовлетворяющий условию обратимости

$$\xi\left(\alpha^i\right)\psi_{[1]}\left(\alpha^j|\alpha^i\right) = \xi\left(\alpha^j\right)\psi_{[1]}\left(\alpha^i|\alpha^j\right).$$

В работе [13] доказано, что для любой матрицы переходных вероятностей $\mathbf{\Psi}_{[s]} = [\underbrace{\mathbf{\Psi}_{[1]} \times \cdots \times \mathbf{\Psi}_{[1]}}_{s}]$, соответствующей разреженной марковской цепи (т. е. большему шагу эволюции), построенные на их основе меры сходства

$$\varkappa_s\left(\alpha^i,\alpha^j\right)=\frac{\psi_{[s]}\left(\alpha^i|\alpha^j\right)}{\xi\left(\alpha^i\right)}$$

обладают свойствами потенциальной функции на множестве аминокислот, образуя неотрицательно определенную матрицу значений парного сходства для любого $s>0$ и погружая множество аминокислот в гипотетическое линейное пространство $\tilde{A}\subset A$ с евклидовой метрикой [14]

$$\rho(\alpha^i,\alpha^j)=\left(\varkappa(\alpha^i,\alpha^i)+\varkappa(\alpha^j,\alpha^j)-2\varkappa(\alpha^i,\alpha^j)\right)^{1/2}, \tag{1}$$

где $\varkappa(\alpha^i,\alpha^j)$ — любая из функций $\varkappa_s(\alpha^i,\alpha^j),\quad s=1,2,\ldots$

Именно эту метрику предлагается использовать в данной работе для сравнения аминокислот.

Аналогичным образом может быть введена метрика на множестве нуклеотидов, составляющих нуклеотидные последовательности.

3 Метрики на основе оптимального выравнивания символьных последовательностей

3.1 Выравнивание символьных последовательностей и расширенные последовательности

Пусть Ω — множество всех последовательностей над некоторым конечным алфавитом $A=\{\alpha^1,\ldots,\alpha^m\}$, в частности алфавитом двадцати аминокислот или четырех нуклеотидов. И пусть $\omega'=(\alpha'_1,\alpha'_2,\ldots,\alpha'_{N'})\in\Omega$ и $\omega'=(\alpha''_1,\alpha''_2,\ldots,\alpha''_{N''})\in\Omega$ — две конкретные последовательности длины N' и N'' соответственно, состоящие из элементов $\alpha'_i,\alpha''_j\in A$, $i=1,\ldots,N'$, $j=1,\ldots,N''$.

Пусть также определена метрика на множестве элементов последовательностей, например в соответствии с (1), обладающая согласно определению следующими свойствами:

$$\left.\begin{array}{l}\rho(\alpha',\alpha''):A\times A\to R;\\ \rho(\alpha',\alpha'')\geqslant 0\ \forall\ \alpha',\alpha''\in A;\\ \rho(\alpha,\alpha)=0\ \forall\ \alpha\in A;\\ \rho(\alpha',\alpha'')+\rho(\alpha'',\alpha''')\geqslant\ \rho(\alpha',\alpha'''),\ \forall\ \alpha',\ \alpha'',\alpha'''\in A.\end{array}\right\} \tag{2}$$

Под выравниванием двух последовательностей $\omega'=(\alpha'_1,\alpha'_2,\ldots,\alpha'_{N'})\in\Omega$ длины N' и $\omega'=(\alpha''_1,\alpha''_2,\ldots,\alpha''_{N''})\in\Omega$ длины N'' понимается приведение их к одинаковой длине путем вставок так называемых «пропусков» в некоторые позиции последовательностей. Пример парного выравнивания двух последовательностей приведен на рис. 2.

α'_1	α'_2	–	α'_3	α'_4	α'_5	α'_6	α'_7
	|		|	|		|
–	α''_1	α''_2	α''_3	α''_4	–	–	α''_5

Рис. 2 Пример парного выравнивания двух последовательностей

Выравнивание естественно представить в виде таблицы парных соответствий элементов сравниваемых последовательностей, в которой пропускам соответствуют нули:

$$\mathbf{w}:\quad\left\{\begin{array}{cccccccc}\mathbf{w}_1&\mathbf{w}_2&\mathbf{w}_3&\mathbf{w}_4&\mathbf{w}_5&\mathbf{w}_6&\mathbf{w}_7&\mathbf{w}_8\\ 1&2&0&3&4&5&6&7\\ 0&1&2&3&4&0&0&5\end{array}\right..$$

Число столбцов данной таблицы определяет длину выравнивания, которая для рассматриваемого примера составляет $N_{\mathbf{w}} = 8$.

Будем называть выравнивание $\mathbf{w}$ допустимым, если оно не содержит два пропуска в одной позиции: $I_{\mathbf{w}} = \{i : \mathbf{w}_{i,1} = \mathbf{w}_{i,2} = 0\} = \varnothing$.

Множество всех допустимых выравниваний пары последовательностей длин N' и N'' будем обозначать $W_{N'N''}$.

Далее будем рассматривать только допустимые выравнивания.

Последовательность $\tilde{\omega}'$, полученную из исходной последовательности ω' путем вставки в нее пропусков, будем называть расширенной последовательностью. Символом $\tilde{\Omega}$ обозначим множество всех возможных расширенных последовательностей над расширенным алфавитом $\tilde{A} = A \cup \{-\} = \{\alpha^1, \ldots, \alpha^m, -\} = \{\tilde{\alpha}^1, \ldots, \tilde{\alpha}^m, \tilde{\alpha}^{m+1}\}$.

В результате конкретного выравнивания $\mathbf{w} = \mathbf{w}(\omega', \omega'')$ пары последовательностей $\omega' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_{N'}) \in \Omega$ и $\omega' = (\alpha''_1, \alpha''_2, \ldots, \alpha''_{N''}) \in \Omega$ образуются расширенные последовательности $\tilde{\omega}' = (\tilde{\alpha}'_1, \tilde{\alpha}'_2, \ldots, \tilde{\alpha}'_{N_{\mathbf{w}}}) \in \tilde{\Omega}$ и $\omega'' = (\tilde{\alpha}''_1, \tilde{\alpha}''_2, \ldots, \tilde{\alpha}''_{N_{\mathbf{w}}}) \in \tilde{\Omega}$ одинаковой длины $N_{\mathbf{w}}$, причем

$$\tilde{\alpha}'_{\mathbf{w},i} = \begin{cases} \alpha'_{\mathbf{w}_{i,1}}, & \mathbf{w}_{i,1} \neq 0 \\ -, & \mathbf{w}_{i,1} = 0 \end{cases} ; \quad \tilde{\alpha}''_{\mathbf{w},i} = \begin{cases} \alpha''_{\mathbf{w}_{i,2}}, & \mathbf{w}_{i,2} \neq 0 \\ -, & \mathbf{w}_{i,2} = 0 \end{cases} , \; i = 1, \ldots, N_{\mathbf{w}}. \tag{3}$$

3.2 Метрика на расширенном множестве элементов последовательностей

Продолжим функцию $\rho(\alpha', \alpha'')$ на расширенное множество $\tilde{A} = A \cup \{-\} = \{\alpha^1, \ldots, \alpha^m, -\} = \{\tilde{\alpha}^1, \ldots, \tilde{\alpha}^m, \tilde{\alpha}^{m+1}\}$ элементов последовательностей следующим образом:

$$\tilde{\rho}(\alpha', \alpha'') = \rho(\alpha', \alpha'') \; \forall \; \alpha', \alpha'' \in A; \quad \tilde{\rho}(-, -) = 0. \tag{4}$$

В дополнение к (4) необходимо также определить значения несходства элементов исходного множества с пропуском $\tilde{\rho}(\alpha, -), \; \alpha \in A$, которые в общем случае могут быть различны.

Теорема 1. *Для того чтобы функция $\tilde{\rho}(\alpha', \alpha'')$, определенная согласно (4), являлась метрикой на расширенном множестве элементов, достаточно, чтобы выполнялось условие:*

$$\tilde{\rho}(\alpha, -) \geqslant const = \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \; \forall \; \alpha \in A.$$

Доказательство теоремы приведено в приложении 1.

3.3 Меры несходства последовательностей, условные относительно выравнивания

Для фиксированного допустимого выравнивания $\mathbf{w}(\omega', \omega'') \in W_{N'N''}$ меры несходства пары последовательностей ω' и ω', условные относительно данного выравнивания, определим двумя способами:

$$r_1(\omega', \omega''|\mathbf{w}) = \sum_{i: \mathbf{w}_{i,1} \neq 0, \mathbf{w}_{i,2} \neq 0} \rho(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) + \sum_{i: \; \mathbf{w}_{i,1}=0 \text{ или } \mathbf{w}_{i,2}=0} \beta;$$

$$r_2(\omega', \omega''|\mathbf{w}) = \sqrt{\sum_{i: \mathbf{w}_{i,1} \neq 0, \mathbf{w}_{i,2} \neq 0} \rho^2(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) + \sum_{i: \; \mathbf{w}_{i,1}=0 \text{ или } \mathbf{w}_{i,2}=0} \beta^2},$$

где

$$\beta = \tilde{\rho}(\alpha, -) \geqslant \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \; \forall \; \alpha \in A. \tag{5}$$

В терминах расширенных последовательностей, с учетом соответствий элементов исходных и расширенных последовательностей (3), определяемых выравниванием $\mathbf{w}$, приведенные меры несходства могут быть записаны в более компактной форме:

$$r_1(\omega', \omega''|\mathbf{w}) = \sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ ); \tag{6}$$

$$r_2(\omega', \omega''|\mathbf{w}) = \sqrt{\sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ )}. \tag{7}$$

3.4 Метрики на множестве последовательностей на основе оптимального выравнивания

Определим две меры несходства последовательностей на основе (6) и (7) соответственно:

$$\mathrm{r}_1(\omega', \omega'') = \min_{\mathbf{w}\in W_{N'N''}} r_1(\omega', \omega''|\mathbf{w}) = \min_{\mathbf{w}\in W_{N'N''}} \sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ ); \tag{8}$$

$$\mathrm{r}_2(\omega', \omega'') = \min_{\mathbf{w}\in W_{N'N''}} r_2(\omega', \omega''|\mathbf{w}) = \min_{\mathbf{w}\in W_{N'N''}} \sqrt{\sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ )}. \tag{9}$$

Теорема 2. *Для любой метрики $\tilde{\rho}(\tilde{\alpha}', \tilde{\alpha}'')$, $\tilde{\alpha}', \tilde{\alpha}'' \in \tilde{A}$, на расширенном множестве элементов $\tilde{A} = A \cup \{-\} = \{\alpha^1, ..., \alpha^m, -\} = \{\tilde{\alpha}^1, ..., \tilde{\alpha}^m, \tilde{\alpha}^{m+1}\}$ меры несходства последовательностей* (8) *и* (9) *обладают свойствами метрики.*

Доказательство теоремы приведено в приложении 2.

4 Алгоритм вычисления метрики на множестве последовательностей

Критерии (8) и (9) относятся к классу парно-сепарабельных целевых функций, поскольку состоят из слагаемых, каждое из которых зависит только от двух соседних переменных. Минимум таких целевых функций может быть найден при помощи процедуры динамического программирования, аналогичной процедуре Нидлмана–Вунша для поиска оптимального глобального выравнивания последовательностей, максимизирующей их сходство [4].

Идея алгоритма заключается в рекуррентном вычислении неизвестных значений несходства $F_{i,j}$ начальных фрагментов последовательностей $(\alpha'_1, \alpha'_2, \ldots, \alpha'_i)$ и $(\alpha''_1, \alpha''_2, \ldots, \alpha''_j)$ на основе уже найденных значений. Для критерия (8) вычисление осуществляется по формуле:

$$F_{i,j} = \min \begin{cases} F_{i-1,j-1} + \rho^2(\alpha'_i, \alpha''_j); \\ F_{i-1,j} + \beta; \\ F_{i,j-1} + \beta, \end{cases}$$

а для критерия (9) — по формуле

$$F_{i,j} = \min \begin{cases} F_{i-1,j-1} + \rho(\alpha'_i, \alpha''_j); \\ F_{i-1,j} + \beta; \\ F_{i,j-1} + \beta. \end{cases}$$

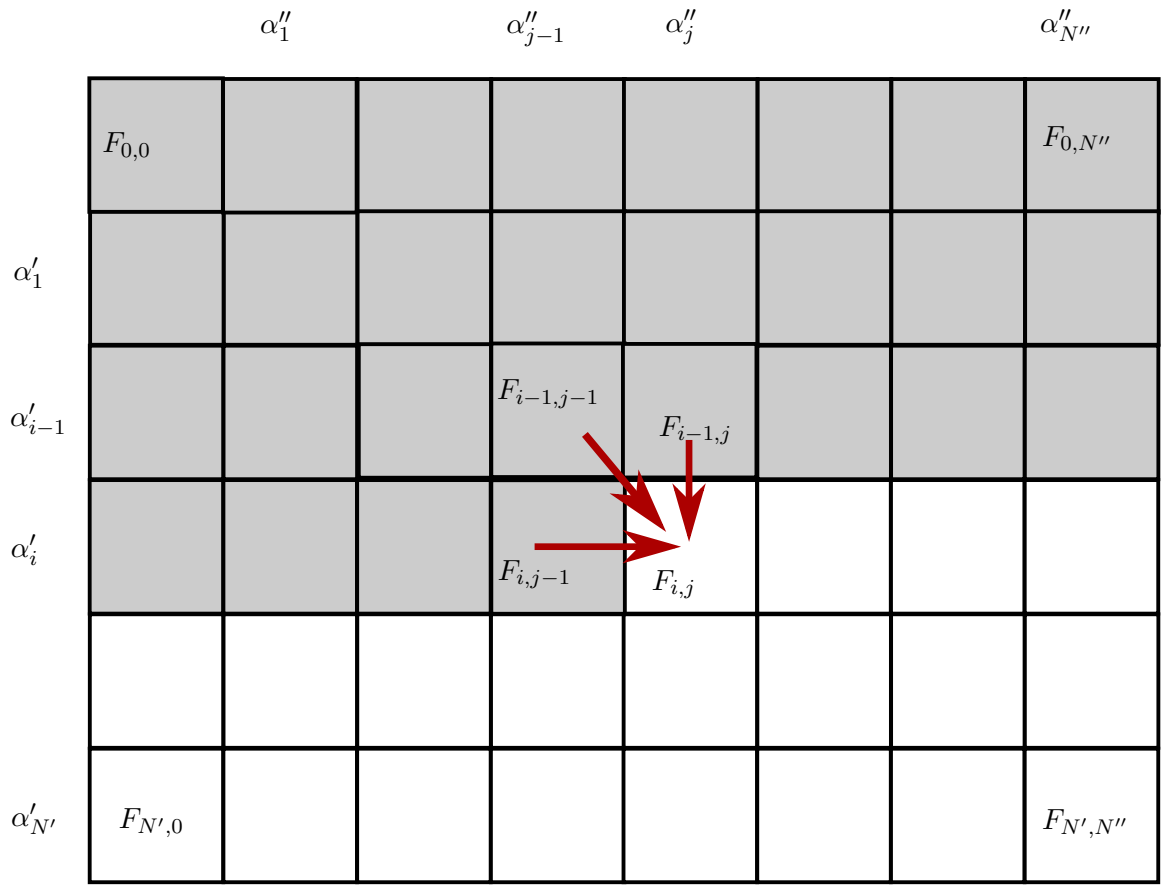


Рис. 3 Схема вычислительного процесса

Для обоих критериев вычисления начинаются с инициализации:

$$F_{0,0} = 0; \quad F_{i,0} = i\beta,\ i = 1, \ldots, N'; \quad F_{0,j} = j\beta,\ j = 1, \ldots, N'',$$

а заканчиваются при достижении концов последовательностей: $r_1(\omega', \omega'') = F_{N'N''}$ и $r_2(\omega', \omega'') = \sqrt{F_{N'N''}}$.

Вычислительный процесс такого вида удобно представлять при помощи таблицы парных соответствий (рис. 3).

Вычислительный процесс заключается в последовательном прохождении всех ячеек таблицы, начиная с левой верхней и заканчивая правой нижней, осуществляя рекуррентные вычисления неполных значений несходства $F_{i,j}$, выбирая и запоминая оптимальное перемещение в соответствующую ячейку (по горизонтали, по вертикали или по диагонали). При этом перемещение по горизонтали соответствует вставке пропуска в последовательность ω', перемещение по вертикали — вставке пропуска в последовательность ω'' и продвижение по диагонали — сравнению элементов последовательностей, стоящих на пересечении соответствующих строки и столбца.

Запомненные для каждой ячейки направления оптимальных перемещений могут быть использованы на обратном ходе процедуры, начинаемом с последней ячейки с координатами (N', N''), для восстановления оптимального пути, однозначно определяющего выравнивание пары последовательностей.

5 Экспериментальное исследование

5.1 Исходные данные

В качестве базы для экспериментального исследования использовались аминокислотные последовательности вирусов простого герпеса из базы данных VIDA (Virus Database at University College London) [36], разделенные специалистами в области молекулярной биологии на три класса на основе лабораторного анализа эволюции вирусов герпеса [37]. Структура исходных данных представлена в табл. 1.

Белки всех трех классов выполняют одну и ту же функцию «Мембранный гликопротеин» (Membrane Glicoprotein), но отличаются друг от друга типом гликопротеинов (например, белки класса 1 являются гликопротеинами-H, а белки класса 2 — гликопротеинами-L). Каждый из рассматриваемых классов включает в себя несколько семейств гомологичных белков (Homologous Protein Families — HPF). Согласно исследованию, проведенному в работе [37], семейства, объединенные в один класс, имеют общего прародителя.

Таблица 1 Исходные данные для анализа последовательностей

Класс	Описание	Гомологические семейства (HPF)	Число белков
1 (109 белков)	Гликопротеин H (glycoprotein H)	12	52
		42	39
		531	18
2 (77 белков)	Гликопротеин L (glycoprotein L)	47	30
		50	32
		114	13
		296	2
3 (48 белков)	Гликопротеин M (glycoprotein M)	20	48

5.2 Сравнение оптимальных выравниваний

Предложенный подход, как и традиционный для молекулярной биологии алгоритм Нидлмана–Вунша, основывается на поиске оптимального глобального парного выравнивания сравниваемых последовательностей.

Найденные оптимальные выравнивания зависят от способа сравнения элементов последовательностей и от используемого значения штрафа на пропуск элементов последовательностей.

Экспериментальное исследование показывает, что при стандартных (использующихся по умолчанию) настройках алгоритма Нидлмана–Вунша и описанном способе построения метрики на множестве биомолекулярных последовательностей со значением штрафа (5) найденные оптимальные выравнивания, как правило, оказываются полностью идентичными либо очень близкими для эволюционно близких последовательностей. С увеличением эволюционного расстояния (увеличением доли точечных мутаций аминокислот) локальные различия между найденными оптимальными выравниваниями могут увеличиваться, однако такие традиционные характеристики качества вырванивания, как общее количество пар сопоставленных друг другу в результате одинаковых (identities) и близких (positives) аминокислот оказываются достаточно близки, что подтверждает осмысленность выравниваний, найденных при помощи предложенного подхода. Примеры оптимальных выравниваний представлены на рис. 4.

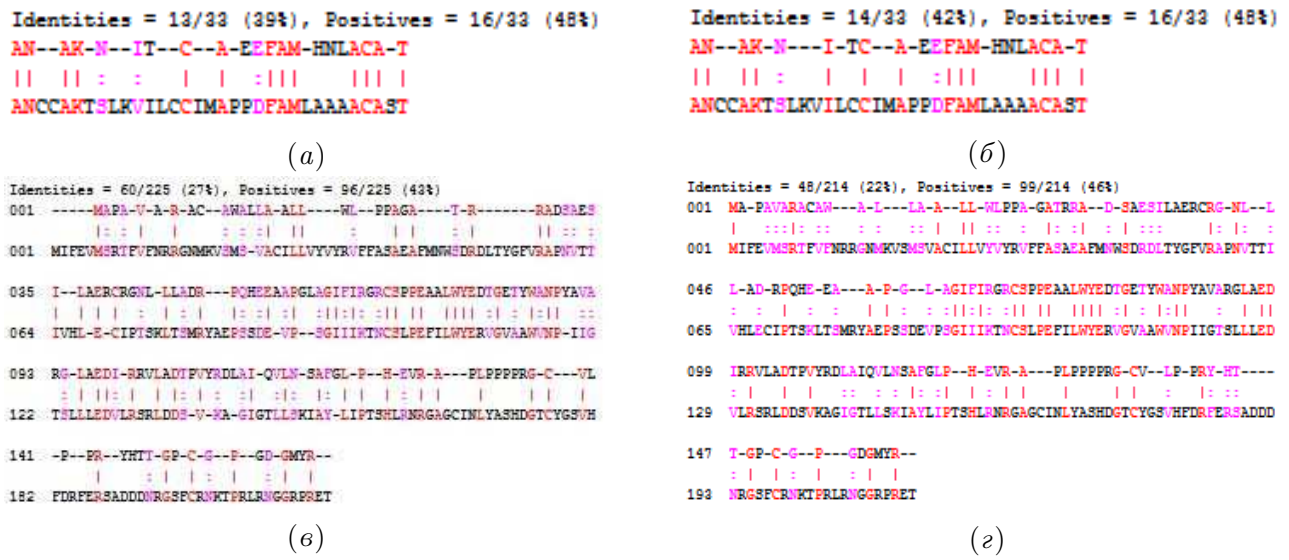


Рис. 4 Примеры оптимальных выравниваний двух пар аминокислотных последовательностей. Выравнивания найдены при помощи предложенного метода (*а* и *в*) и алгоритма Нидлмана–Вунша (*б* и *г*)

5.3 Классификация мембранных гликопротеинов

В данной работе используются 3 базовых способа сравнения последовательностей:

1) мера сходства Нидлмана–Вунша (NW) — $S_1(\omega', \omega'')$;
2) мера сходства Смита–Ватермана (SW) — $S_2(\omega', \omega'')$;
3) предложенная метрика (Metric) — $r(\omega', \omega'')$.

Для обеспечения возможности использования мер сходства Нидлмана–Вунша и Смита–Ватермана совместно с методом опорных векторов (SVM) для каждой из них был выполнен переход в пространство вторичных признаков:

$$K_i(\omega', \omega'') = \left[k_{lt} = \left(S_i^{\langle l \rangle} \right)^T S_i^{\langle t \rangle}, \quad l, t = 1, \ldots, N \right], \quad i = 1, 2.$$

Что касается предложенной метрики, то оказалось, что на рассматриваемом множестве аминокислотных последовательностей она является евклидовой, что позволяет использовать радиальную функцию вида $K_3(\omega', \omega'') = \exp(-\alpha r^2(\omega', \omega''))$. Значение параметра α во всех проведенных экспериментах было установлено равным 0,01.

Следует обратить внимание, что в общем случае предложенный способ построения метрики не гарантирует наличие свойства евклидовости и, соответственно, преобразование $\exp(-\alpha r^2(\omega', \omega''))$ может привести к наличию отрицательных собственных чисел. Однако на практике для $r(\omega', \omega'') = r_2(\omega', \omega'')$ свойство евклидовости обычно выполняется.

Для каждого из трех указанных способов сравнения решались задачи обучения двухклассовому распознаванию каждого из классов (1, 2 и 3) от оставшихся и каждого семейства (hpf 12, 20, 42, 47, 50, 114, 531) от оставшихся, а также задачи обучения попарному распознаванию классов и семейств из табл. 1. Обучение проводилось при помощи SVM.

Качество построенных решающих правил оценивалось по скользящему контролю.

В табл. 2 и 3 приведены проценты ошибок, полученных на скользящем контроле, для случаев, в которых хотя бы два из трех способов сравнения дали различный результат. Жирным выделен лучший результат в каждой строке.

Таблица 2 Проценты ошибок на скользящем контроле при распознавании «один против всех»

Задача	NW	SW	Metric
hpf 12	15,0215	15,0215	**14,5923**
hpf 20	0,4292	**0**	**0**
hpf 42	**0**	0,4292	0,4292
hpf 47	4,721	**0**	**0**
hpf 50	0,4292	**0**	**0**
hpf 114	4,721	0,8584	**0,4292**
hpf 531	**15,0125**	**15,0125**	18,4549
класс 1	0,8584	**0,4292**	**0,4292**
класс 2	0,8584	**0,4292**	**0,4292**
класс 3	0,4292	**0**	**0**

Таблица 3 Проценты ошибок на скользящем контроле при попарном распознавании классов и семейств hpf

Задача	NW	SW	Metric
класс 2 vs класс 3	12,3256	**0**	**0**
hpf 42 vs hpf 47	0,4292	**0**	**0**
hpf 42 vs hpf 114	**0**	1,9231	**0**
hpf 47 vs hpf 114	2,3256	**0**	**0**
hpf 531 vs hpf 12	**48,5714**	51,4286	50,000
hpf 531 vs hpf 42	**1,7544**	3,5088	**1,7544**

Как видно из табл. 2 и 3, предложенный способ сравнения позволяет в большинстве случаев получить наилучший результат, в остальных случаях — близкий к наилучшему, что говорит о его адекватности прикладным задачам анализа аминокислотных последовательностей. Кроме того, очень важным достоинством данного способа сравнения является то, что его применение совместно с SVM требует запоминания и использования при распознавании лишь небольшого количества опорных объектов, а не всех объектов обучающей совокупности, как при использовании традиционных мер сходства Нидлмана–Вунша и Смита–Ватермана, что делает его более эффективным по памяти и скорости работы.

6 Заключение

В данной работе предложен простой способ построения метрики на множестве биомолекулярных последовательностей, основанный, как и традиционные методы, на поиске оптимального парного выравнивания. При соблюдении необременительного условия относительно величины штрафа на пропуск элементов при выравнивании получаемая в результате мера несходства гарантированно обладает свойствами метрики. Доказательства соответствующих теорем приведены в данной статье. Кроме того, в ряде практических случаев (для конкретных конечных множеств последовательностей) данная мера несходства может обладать и свойствами евклидовой метрики, что делает ее применение в сочетании с SVM особенно удобным.

Экспериментальное исследование показало адекватность данного способа сравнения прикладным задачам распознавания аминокислотных последовательностей. Кроме того, очень важным достоинством использования метрики в качестве способа сравнения является то, что его применение совместно с SVM требует запоминания и использования при распознавании лишь небольшого количества опорных объектов, а не всех объектов обучающей совокупности, как при использовании традиционных мер сходства Нидлмана–Вунша и Смита–Ватермана, что делает его более эффективным по памяти и скорости работы.

Приложение 1

Доказательство теоремы 1

Согласно определению (4) и условию теоремы 1 условие неотрицательности $\tilde{\rho}(\alpha', \alpha'') \geqslant 0$ выполняется для всех элементов из расширенного множества $\alpha', \alpha'' \in \tilde{A}$.

Покажем, что для любой тройки элементов $\alpha',\ \alpha'', \alpha''' \in \tilde{A}$ выполняется неравенство треугольника

$$\tilde{\rho}(\alpha', \alpha'') + \tilde{\rho}(\alpha'', \alpha''') \geqslant\ \tilde{\rho}(\alpha', \alpha''')\ \forall\ \alpha',\ \alpha'', \alpha''' \in \tilde{A}. \tag{10}$$

Рассмотрим все возможные способы вхождения пропуска «$-$» в тройку элементов $\alpha', \alpha'', \alpha''' \in \in \tilde{A}$ и покажем, что для каждого из этих случаевнеравенство треугольника выполняется:

а) α'''«$-$», $\alpha', \alpha'' \neq$«$-$».

Неравенство треугольника (10) в данном случае принимает вид:

$$\tilde{\rho}(\alpha', \alpha'') + \tilde{\rho}(\alpha'', -) \geqslant\ \tilde{\rho}(\alpha', -)\ \forall\ \alpha',\ \alpha'' \in A.$$

Подставив $\tilde{\rho}(\alpha', -) = \tilde{\rho}(\alpha'', -) = (1/2) \max\limits_{\eta', \eta'' \in A} \rho(\eta', \eta'')$, получим:

$$\tilde{\rho}(\alpha', \alpha'') + \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \geqslant\ \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'').$$

Очевидно, что это условие выполняется для любых $\alpha',\ \alpha'' \in A$, поскольку $\tilde{\rho}(\alpha', \alpha'')\ \geqslant\ 0$;

б) $\alpha' =$«$-$», $\alpha'', \alpha''' \neq$«$-$». Этот случай абсолютно аналогичен предыдущему. Выполняя аналогичную подстановку $\tilde{\rho}(\alpha'', -) = \tilde{\rho}(\alpha''', -) = (1/2) \max\limits_{\eta', \eta'' \in A} \rho(\eta', \eta'')$ в неравенство треугольника, получим:

$$\frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') + \tilde{\rho}(\alpha'', \alpha''') \geqslant \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'').$$

Соответственно, поскольку $\tilde{\rho}(\alpha'', \alpha''') \geqslant 0\ \forall\ \alpha'', \alpha''' \in \tilde{A}$, в данном случае неравенство треугольника тоже выполняется;

в) $\alpha''\ =$«$-$», $\alpha', \alpha'''\ \neq$«$-$». Неравенство треугольника (10) после подстановки $\tilde{\rho}(\alpha', -)\ = \\ = \tilde{\rho}(\alpha''', -) = (1/2) \max\limits_{\eta', \eta'' \in A} \rho(\eta', \eta'')$ примет вид:

$$\frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') + \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \geqslant\ \tilde{\rho}(\alpha', \alpha''')\ \forall\ \alpha',\ \alpha''' \in A$$

или

$$\max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \geqslant\ \tilde{\rho}(\alpha', \alpha''')\ \forall\ \alpha',\ \alpha''' \in A.$$

Очевидно, что данное неравенство всегда является верным;

г) $\alpha' = \alpha'' =$«$-$», $\alpha''' \neq$«$-$».

В данном случае неравенство треугольника

$$\tilde{\rho}(-,-) + \tilde{\rho}(-,\alpha''') \geqslant \tilde{\rho}(-,\alpha''') \; \forall \; \alpha',\; \alpha'',\alpha''' \in \tilde{A}$$

вырождается в верное равенство:

$$0 + \frac{1}{2} \max_{\eta',\eta'' \in A} \rho(\eta',\eta'') = \frac{1}{2} \max_{\eta',\eta'' \in A} \rho(\eta',\eta'');$$

д) $\alpha' = \alpha'' = \alpha''' =$«$-$». Данный случай является тривиальным, приводя после подстановки к тождеству $0 + 0 = 0$;

е) $\alpha' \neq$«$-$», $\alpha'' \neq$«$-$», $\alpha''' \neq$«$-$».

В данном случае все три элемента принадлежат исходному множеству элементов последовательностей $\alpha',\; \alpha'',\; \alpha''' \in A$, на котором определена метрика $\rho(\alpha',\alpha'') : A \times A \to R$ согласно (2) и, поскольку, согласно (4) $\tilde{\rho}(\alpha',\alpha'') = \rho(\alpha',\alpha'') \; \forall \; \alpha',\alpha'' \in A$, то неравенство треугольника в данном случае тоже выполняется.

Таким образом, для любой тройки элементов $\alpha',\; \alpha'',\alpha''' \in \tilde{A}$ неравенство треугольника выполняется.

Теорема доказана.

Приложение 2

Доказательство теоремы 2

Для доказательства выполнения неравенства треугольника для мер несходства (8) и (9) рассмотрим три последовательности $\omega' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_{N'}) \in \Omega,\; \omega'' = (\alpha''_1, \alpha''_2, \ldots, \alpha''_{N''}) \in \Omega$ и $\omega''' = (\alpha'''_1, \alpha'''_2, \ldots, \alpha'''_{N'''}) \in \Omega$.

Пусть $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega',\omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'',\omega''')$ — два оптимальных выравнивания соответствующих пар последовательностей, например следующих:

$$\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega',\omega'') : \quad \left\{ \begin{array}{cccccc} \hat{\mathbf{w}}_1^{1,2} & \hat{\mathbf{w}}_2^{1,2} & \hat{\mathbf{w}}_3^{1,2} & \hat{\mathbf{w}}_4^{1,2} & \hat{\mathbf{w}}_5^{1,2} & \hat{\mathbf{w}}_6^{1,2} \\ \alpha'_1 & \alpha'_2 & - & \alpha'_3 & - & \alpha'_4 \\ -\alpha''_1 & - & \alpha''_2 & \alpha''_3 & \alpha''_4 & \alpha''_5 \end{array} \right. ;$$

$$\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'',\omega''') : \quad \left\{ \begin{array}{cccccc} \hat{\mathbf{w}}_1^{2,3} & \hat{\mathbf{w}}_2^{2,3} & \hat{\mathbf{w}}_3^{2,3} & \hat{\mathbf{w}}_4^{2,3} & \hat{\mathbf{w}}_5^{2,3} & \hat{\mathbf{w}}_6^{2,3} \\ \alpha''_1 & \alpha''_2 & \alpha''_3 & \alpha''_4 & \alpha''_5 & - \\ - & \alpha'''_1 & \alpha'''_2 & - & \alpha'''_3 & \alpha'''_4 \end{array} \right. .$$

Очевидно, что два таких выравнивания однозначно определяют третье выравнивание $\mathbf{w}^{1,3} = \mathbf{w}(\omega',\omega''')$, сопоставляющее элементы последовательностей ω' и ω''', а также выравнивание элементов всех трех последовательностей сразу $\mathbf{w}^{1,2,3} = \mathbf{w}(\omega',\omega'',\omega''')$:

$$\mathbf{w}^{1,3} = \mathbf{w}(\omega',\omega''') : \quad \left\{ \begin{array}{cccccc} \mathbf{w}_1^{1,3} & \mathbf{w}_2^{1,3} & \mathbf{w}_3^{1,3} & \mathbf{w}_4^{1,3} & \mathbf{w}_5^{1,3} & \mathbf{w}_6^{1,3} \\ \alpha'_1 & \alpha'_2 & - & \alpha'_3 & \alpha'_4 & - \\ - & - & \alpha'''_1 & \alpha'''_2 & \alpha'''_3 & \alpha'''_4 \end{array} \right. ;$$

$$\mathbf{w}^{1,2,3} = \mathbf{w}(\omega',\omega'',\omega''') : \quad \left\{ \begin{array}{ccccccc} \mathbf{w}_1^{1,2,3} & \mathbf{w}_2^{1,2,3} & \mathbf{w}_3^{1,2,3} & \mathbf{w}_4^{1,2,3} & \mathbf{w}_5^{1,2,3} & \mathbf{w}_6^{1,2,3} & \mathbf{w}_7^{1,2,3} \\ \alpha'_1 & \alpha'_2 & - & \alpha'_3 & - & \alpha'_4 & - \\ \alpha''_1 & - & \alpha''_2 & \alpha''_3 & \alpha''_4 & \alpha''_5 & - \\ - & - & \alpha'''_1 & \alpha'''_2 & - & \alpha'''_3 & \alpha'''_4 \end{array} \right. .$$

Следует обратить внимание, что выравнивания $\mathbf{w}^{1,3}$ и $\mathbf{w}^{1,2,3}$ в отличие от $\hat{\mathbf{w}}^{1,2}$ и $\hat{\mathbf{w}}^{2,3}$ в общем случае не являются оптимальными.

Каждое из рассмотренных выравниваний порождает свои расширенные последовательности в соответствии с (3):

$$
\begin{array}{ll}
\hat{\mathbf{w}}^{1,2}: & \tilde{\omega}'(\hat{\mathbf{w}}^{1,2}) = \{\tilde{\alpha}'_{\hat{\mathbf{w}}^{1,2},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{1,2}}\}, \quad \tilde{\omega}''(\hat{\mathbf{w}}^{1,2}) = \{\tilde{\alpha}''_{\hat{\mathbf{w}}^{1,2},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{1,2}}\};\\
\hat{\mathbf{w}}^{2,3}: & \tilde{\omega}''(\hat{\mathbf{w}}^{2,3}) = \{\tilde{\alpha}''_{\hat{\mathbf{w}}^{2,3},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{2,3}}\}, \quad \tilde{\omega}'''(\hat{\mathbf{w}}^{2,3}) = \{\tilde{\alpha}'''_{\hat{\mathbf{w}}^{2,3},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{2,3}}\};\\
\mathbf{w}^{1,3}: & \tilde{\omega}'(\mathbf{w}^{1,3}) = \{\tilde{\alpha}'_{\mathbf{w}^{1,3},i}, i = 1,\ldots,N_{\mathbf{w}^{1,3}}\}, \quad \tilde{\omega}'''(\mathbf{w}^{1,3}) = \{\tilde{\alpha}'''_{\mathbf{w}^{1,3},i}, i = 1,\ldots,N_{\mathbf{w}^{1,3}}\};\\
\mathbf{w}^{1,2,3}: & \left\{\begin{array}{l}
\tilde{\omega}'(\mathbf{w}^{1,2,3}) = \{\tilde{\alpha}'_{\mathbf{w}^{1,2,3},i}\ ,\ i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}\},\\
\tilde{\omega}''(\mathbf{w}^{1,2,3}) = \{\tilde{\alpha}''_{\mathbf{w}^{1,2,3},i}\ ,\ i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}\},\\
\tilde{\omega}'''(\mathbf{w}^{1,2,3}) = \{\tilde{\alpha}'''_{\mathbf{w}^{1,2,3},i}\ ,\ i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}\}.
\end{array}\right.
\end{array}
$$

Для каждой пары последовательностей (ω',ω''),$(\omega'',\omega]''$) и (ω',ω''') рассмотрим векторы одинаковой длины $N_{\mathbf{w}^{1,2,3}}$, составленные из значений метрики (4) между их элементами, поставленными в соответствие выравниванием $\mathbf{w}^{1,2,3}$:

$$
\begin{aligned}
\mathbf{x}^{1,2} &= \mathbf{r}_1(\omega',\omega''|\mathbf{w}^{1,2,3}) = [\tilde{\rho}\left(\tilde{\alpha}'_{\mathbf{w}^{1,2,3},i}, \tilde{\alpha}''_{\mathbf{w}^{1,2,3},i}\right), \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}]^{\mathrm{T}};\\
\mathbf{x}^{2,3} &= \mathbf{r}_1(\omega'',\omega'''|\mathbf{w}^{1,2,3}) = [\tilde{\rho}\left(\tilde{\alpha}''_{\mathbf{w}^{1,2,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,2,3},i}\right), \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}]^{\mathrm{T}};\\
\mathbf{x}^{1,3} &= \mathbf{r}_1(\omega',\omega'''|\mathbf{w}^{1,2,3}) = [\tilde{\rho}\left(\tilde{\alpha}'_{\mathbf{w}^{1,2,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,2,3},i}\right), \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}]^{\mathrm{T}}.
\end{aligned}
$$

Заметим, что, согласно свойству метрики (2), для каждой тройки $i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}$ элементов этих векторов справедливы неравенства:

$$
x_i^{1,2} + x_i^{2,3} \geqslant x_i^{1,3}, \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}.
$$

При этом очевидно, что также будет справедливо неравенство:

$$
\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,2} + \sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{2,3} \geqslant \sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,3}. \tag{11}
$$

Кроме того, нетрудно убедиться, что в этом случае евклидова норма вектора $\mathbf{x}^{1,3}$ не может превосходить сумму евклидовых норм векторов $\mathbf{x}^{1,2}$ и $\mathbf{x}^{2,3}$:

$$
\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} \left(x_i^{1,2}\right)^2} + \sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} \left(x_i^{2,3}\right)^2} \geqslant \sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} \left(x_i^{1,3}\right)^2}. \tag{12}
$$

Заметим, что значения $\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,2}$ и $\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{2,3}$, входящие в неравенство (11), равны, соответственно, несходству последовательностей (ω',ω'') и (ω'',ω'''), определяемому согласно (8) на основе их оптимальных выравниваний $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega',\omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'',\omega''')$, поскольку $\tilde{\rho}(-,-) = 0$:

$$
\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,2} = \sum_{i=1}^{N_{\hat{\mathbf{w}}^{1,2}}} \tilde{\rho}\left(\tilde{\alpha}'_{\hat{\mathbf{w}}^{1,2},i}, \tilde{\alpha}''_{\hat{\mathbf{w}}^{1,2},i}\right) = r_1(\omega',\omega''|\hat{\mathbf{w}}^{1,2}) = \min_{\mathbf{w}\in W_{N'N''}} r_1(\omega',\omega''|\mathbf{w}) = r_1(\omega',\omega'');
$$

$$
\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{2,3} = \sum_{i=1}^{N_{\hat{\mathbf{w}}^{2,3}}} \tilde{\rho}\left(\tilde{\alpha}''_{\hat{\mathbf{w}}^{2,3},i}, \tilde{\alpha}'''_{\hat{\mathbf{w}}^{2,3},i}\right) = r_1(\omega'',\omega'''|\hat{\mathbf{w}}^{2,3}) = \min_{\mathbf{w}\in W_{N''N'''}} r_1(\omega'',\omega'''|\mathbf{w}) = r_1(\omega'',\omega''').
$$

Аналогично, значения $\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,2}\right)^2}$ и $\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{2,3}\right)^2}$, входящие в неравенство (12), равны, соответственно, несходству последовательностей (ω', ω'') и (ω'', ω'''), определяемому согласно (9) для тех же оптимальных выравниваний $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega', \omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'', \omega''')$:

$$\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,2}\right)^2} = \sqrt{\sum_{i=1}^{N_{\hat{\mathbf{w}}^{1,2}}}\tilde{\rho}^2\left(\tilde{\alpha}'_{\hat{\mathbf{w}}^{1,2},i}, \tilde{\alpha}''_{\hat{\mathbf{w}}^{1,2},i}\right)} = r_2(\omega', \omega''|\hat{\mathbf{w}}^{1,2}) = r_2(\omega', \omega'');$$

$$\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{2,3}\right)^2} = \sqrt{\sum_{i=1}^{N_{\hat{\mathbf{w}}^{2,3}}}\tilde{\rho}^2\left(\tilde{\alpha}''_{\hat{\mathbf{w}}^{2,3},i}, \tilde{\alpha}'''_{\hat{\mathbf{w}}^{2,3},i}\right)} = r_2(\omega'', \omega'''|\hat{\mathbf{w}}^{2,3}) = r_2(\omega'', \omega''').$$

При этом следует заметить, что для пары последовательностей (ω', ω''') оптимальное выравнивание не определено и значения $\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,3}$ и $\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,3}\right)^2}$ для них равны определяемым, соответственно, согласно (6) и (7) значениям несходства, условного относительно выравнивания $\mathbf{w}^{1,3}$, порожденного оптимальными выравниваниями $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega', \omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'', \omega''')$:

$$\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,3} = \sum_{i=1}^{N_{\mathbf{w}^{1,3}}}\tilde{\rho}\left(\tilde{\alpha}'_{\mathbf{w}^{1,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,3},i}\right) = r_1(\omega', \omega'''|\mathbf{w}^{1,3});$$

$$\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,3}\right)^2} = \sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,3}}}\tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w}^{1,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,3},i})} = r_2\left(\omega', \omega'''|\mathbf{w}^{1,3}\right).$$

Таким образом, выполняются неравенства:

$$r_1(\omega', \omega'') + r_1(\omega'', \omega''') \geqslant r_1(\omega', \omega'''|\mathbf{w}^{1,3}),$$

$$r_2(\omega', \omega'') + r_2(\omega'', \omega''') \geqslant r_2(\omega', \omega'''|\mathbf{w}^{1,3}).$$

Более того, данные неравенства останутся верными и в случае, если вместо выравнивания $\mathbf{w}^{1,3} = \mathbf{w}(\omega', \omega''')$, порожденного выравниваниями $\hat{\mathbf{w}}^{1,3} = \hat{\mathbf{w}}(\omega', \omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'', \omega''')$, рассмотреть оптимальное выравнивание $\hat{\mathbf{w}}^{1,3} = \hat{\mathbf{w}}(\omega', \omega''')$ последовательностей ω' и ω'', поскольку оптимальное выравнивание по определению обеспечивает значение несходства, не превышающее значения, вычисленного для любого другого варианта выравнивания последовательностей:

$$r_1(\omega', \omega''') = r_1(\omega', \omega'''|\hat{\mathbf{w}}^{1,3}) \leqslant r_1(\omega', \omega'''|\mathbf{w}^{1,3});$$

$$r_2(\omega', \omega''') = r_2(\omega', \omega'''|\hat{\mathbf{w}}^{1,3}) \leqslant r_2(\omega', \omega'''|\mathbf{w}^{1,3}).$$

Следовательно, для любой тройки последовательностей неравенства треугольников $r_1(\omega', \omega'') + r_1(\omega'', \omega''') \geqslant r_1(\omega', \omega''')$ и $r_2(\omega', \omega'') + r_2(\omega'', \omega''') \geqslant r_2(\omega', \omega''')$ выполняются, и меры несходства, определенные согласно (8) и (9), являются метриками на множестве последовательностей.

Теорема доказана.

Литература

[1] *Needleman S. B., Wunsch C. D.* A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol., 1970. Vol. 48. No. 3. P. 443–453. doi: 10.1016/0022-2836(70)90057-4.

[2] *Smith T. F., Waterman M. S.* Identification of common molecular subsequences // J. Mol. Biol., 1981. Vol. 147. No. 1. P. 195–197. doi: 10.1016/0022-2836(81)90087-5.

[3] *Zhang Z., Schwartz S., Wagner L., Miller W.* A greedy algorithm for aligning DNA sequences // J. Comput. Biol., 2000. No. 7. P. 203–14. doi: 10.1089/10665270050081478.

[4] *Дурбин Р., Эдди Ш., Крог А., Митчисон Г.* Анализ биологических последовательностей / Пер. с англ. — М.: Ижевск, 2006. 480 с. (*Durbin R., Eddy S., Krogh A., and Mitchison G.* Biological sequence analysis: Probabilistic models of proteins and nucleic acids. — Cambridge Univesrity Press, 1998. 356 p.)

[5] *Vapnik V. N.* Statistical learning theory. — Wiley-Interscience, 1998. 768 p.

[6] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B.* Alignment scores in a regularized support vector classification method for fold recognition of remote protein families. — Center for Discrete Mathematics and Theoretical Computer Science. Rutgers University, State University of New Jersey, 2001. 33 p.

[7] *Pekalska E., Paclic P., Duin R.* A generalized kernel approach to dissimilarity-based classification // J. Mach. Learn. Res., 2001. Vol. 2. P. 175–211.

[8] *Liao Li, Noble W. S.* Combining pairwise sequence similarity and support vector machines for remote protein homology detection // 6th Annual Conference (International) on Computational Molecular Biology Proceedings, 2002. P. 225–232.

[9] *Schölkopf B., Tsuda K., Vert J.-P.* Kernel methods in computational biology. — MIT Press, 2004. 410 p.

[10] *Ben-Hur A., Ong C. S., Sonnenburg S., Schölkopf B., Rätsch G.* Support vector machines and kernels for computational biology // PLoS Comput. Biol., 2008. Vol. 4. No. 10. P. 1–10.

[11] *Середин О. С.* Линейные методы распознавания образов на множествах объектов произвольной природы, представленных попарными сравнениями. Общий случай // Известия ТулГУ, Серия Естественные науки. — Тула: Изд-во ТулГУ, 2012. Т. 1. С. 141–152.

[12] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. 384 с.

[13] *Сулимова В. В.* Потенциальные функции для анализа сигналов и символьных последовательностей разной длины. — Тула, 2009. Дисс. ... канд. наук. 122 с.

[14] *Моттль В. В.* Метрические пространства, допускающие введение линейных операций и скалярного произведения // Докл. РАН, 2003. Т. 388. №3. С. 312–315.

[15] *Leslie C., Eskin E., and Noble W.* The spectrum kernel: A string kernel for SVM protein classification // Pacific Symposium on Biocomputing Proceedings, 2002. P. 564–575.

[16] *Qiu J., Hue M., Ben-Hur A., Vert J.-P., Noble W. S.* A structural alignment kernel for protein structures // Bioinformatics, 2007. Vol. 23. №9. P. 1090–1098.

[17] *Sun L., Ji S., Ye J.* Adaptive diffusion kernel learning from biological networks for protein function prediction // BMC Bioinformatics, 2008. Vol. 9. No. 1. P. 1–14. doi: 10.1186/1471-2105-9-162. `http://www.biomedcentral.com/1471-2105/9/162`.

[18] *Miklós I., Novak A., Satija R., Lyngsø R., Hein J.* Stochastic models of sequence evolution including insertion-deletion events // Stat. Methods Med. Res., 2009. Vol. 18. P. 453–485. `http://ramet.elte.hu/~miklosi/StatAlignReview2008.pdf`.

[19] *Onodera T., Shibuya T.* The gapped spectrum kernel for support vector machines // 9th Conference (International) MLDM Proceedings. — Berlin – Heidelberg – New York: Springer, 2013. P. 1–15. doi: 10.1007/978-3-642-39712-7_1.

[20] *Baisero A., Pokorny F. T., Ek C.H.* On a family of decomposable kernels on sequences // CoRR, 2015. arXiv:/1501.06284.

[21] *Seeger M.* Covariance kernels from baye ian generative models // Stat. Methods Med. Res., 2009. Vol. 18. P. 453–485.

[22] *Абрамов В. И., Середин О. С., Сулимова В. В.* Метод опорных объектов для обучения распознаванию образов в евклидовых метрических пространствах // Международная конференция «Интеллектуализация обработки информации» (ИОИ-9). — Черногория, 2012. С. 5–8.

[23] *Абрамов В. И., Середин О. С., Моттль В. В.* Обучение распознаванию образов по методу опорных объектов в евклидовых метрических пространствах с аффинными операциями // Известия ТулГУ, Естественные науки. — Тула: Изд-во ТулГУ, 2013, Вып. 2. Ч. 1. С. 119–136.

[24] *Hein M., Bousquet O., Schölkopf B.* Maximal margin classification for metric spaces // J. Comput. Syst. Sci., 2005. Vol. 71. Iss. 3. P. 333–359.

[25] *Xu W.* Non-euclidean dissimilarity data in pattern recognition. 2012. Ph.D. Thesis.

[26] *Середин О. С., Абрамов В. И., Моттль В. В.* Аффинные операции в псевдоевклидовом линейном пространстве // Известия ТулГУ, Естественные науки. — Тула: Изд-во ТулГУ, 2014. Вып. 3. С. 178–196.

[27] *Hancock E. R., Xu E., Wilson R. C.* Pattern recognition with non-Euclidean similarities // Man–Machine Interactions, 2014. Vol. 3. P. 3–15.

[28] *Середин О. С., Моттль В. В.* Метод опорных объектов для обучения распознаванию образов в произвольных метрических пространствах // Известия ТулГУ, Естественные науки. — Тула: Изд-во ТулГУ, 2015. Вып. 4. С. 49–66.

[29] *Pekalska E. M.* Dissimilarity representations in pattern recognition. Concepts, theory and applications. 2005. PhD Thesis. 344 p.

[30] *Bellet A., Harbrad A., Sebban M.* A survey on metric learning for feature vectors and structured data // CoRR, 2013. abs/1306.6709. http://arxiv.org/abs/1306.6709.

[31] *Wang J., Sun K., Sha F., Marchand-Maillet S., Kalousis K.* Two-stage metric learning // 31st Conference (International) on Machine Learning Proceedings, Cycle 2. — JMLR.org, 2014. Vol. 32. P. 370–378. http://jmlr.org/proceedings/papers/v32/wangc14.html.

[32] *Xing E. P., Ng A. Y., Jordan M. I., Russel S.* Distance metric learning, with application to clustering with side-information // Advances in neural information processing systems 15 / Eds. S. Becker, S. Thrun, K. Obermayer. — MIT Press, 2003. P. 521–528. http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.pdf.

[33] *Schultz M., Joachims T.* Learning a distance metric from relative comparisons // Advances in neural information processing systems 16 / Eds. S. Thrun, L. K. Saul, B. Schölkopf. — MIT Press, 2004. P. 41–48. http://papers.nips.cc/paper/2366-learning-a-distance-metric-from-relative-comparisons.pdf.

[34] *Wang J., Do H., Woznica A., Kalousis A.* Metric learning with multiple kernels // Advances in neural information processing systems 24 / Eds. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett. — MIT Press, 2011. P. 1170–1178. http://papers.nips.cc/paper/4399-metric-learning-with-multiple-kernels.pdf.

[35] *Dayhoff M., Schwarts R., Orcutt B.* A model of evolutionary change in proteins // Atlas of protein sequences and structures. — National Biometrical Research Foundation, 1978. Vol. 5. Suppl. 3. P. 345–352.

[36] Virus Database at University College London (VIDA). http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA3/VIDA.html.

[37] *McGeoch D. J., Rixon F. J., Davison A. J.* Topics in herpesvirus genomics and evolution // Virus Res., 2006. No. 117. P. 90–104. doi: 10.1016/j.virusres.2006.01.002.

Поступила в редакцию 31.08.2016

Metrics on the basis of optimal alignment of biomolecular sequences*

***V. V. Sulimova*[1], *O. S. Seredin*[1], *and V. V. Mottl*[2]**
vsulimova@yandex.ru; oseredin@yandex.ru; vmottl@yandex.ru
[1]Tula State University, 92 Lenina Ave., Tula, Russia
[2]Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

Background: It is important for biomolecular sequences analysis to have an appropriate way for comparing them. From the point of view of advanced methods of data analysis, the most preferred way for comparing objects is a dissimilarity measure, possessing metric's properties. From the other side, from the point of view of the molecular biology, it is important to take into account biological features of the compared objects. Besides, the computational effectiveness and the possibility of further using convenient instruments of data analysis are also important. There are a number of ways for comparing biomolecular sequences, though no one of them possess the all required properties.

Methods: This paper proposes a simple enough way for computing metrics for biomolecular sequences. The proposed approach, following traditional ways for biomolecular sequences comparing, is based on finding an optimal pairwise alignment and on the model of mutual changes of amino acids at the process of evolution.

Concluding Remarks: It is proved that the proposed dissimilarity measure is a metric. So, it can be used at the advanced methods of data analysis, saving computational advantages of support vector machine without introducing features of objects or(and) an inner product. The experimental results confirm usability of the proposed metric for membrane glycoprotein classification.

Keywords: *metric; comparing sequences; optimal sequence alignment; biomolecular sequences*

DOI: 10.21469/22233792.2.3.03

References

[1] Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3):443–453. doi: 10.1016/0022-2836(70)90057-4.

[2] Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147(1):195–197. doi: 10.1016/0022-2836(81)90087-5.

[3] Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7:203–14. doi: 10.1089/10665270050081478.

[4] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press. 356 p.

[5] Vapnik, V. N. *Statistical learning theory*. Wiley-Interscience, 1998. 768 p.

[6] Mottl, V. V., S. D. Dvoenko, O. S. Seredin, C. A. Kulikowski, and I. B. Muchnik. 2001. *Alignment scores in a regularized support vector classification method for fold recognition of remote protein families*. Center for Discrete Mathematics and Theoretical Computer Science. Rutgers University, State University of New Jersey. 33 p.

*The research was supported by the Russian Foundation for Basic Research (grant 15-07-08967).

[7] Pekalska, E., P. Paclic, and R. Duin. 2001. A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.* 2:175–211.

[8] Liao, Li, and W. S. Noble. 2002. Combining pairwise sequence similarity and support vector machines for remote protein homology detection // *6th Annual Conference (International) on Computational Molecular Biology Proceedings*. 225–232.

[9] Schölkopf, B., K. Tsuda, and J.-P.Vert. 2004. *Kernel methods in computational biology*. MIT Press. 410 p.

[10] Ben-Hur, A., C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. 2008. Support vector machines and kernels for computational biology. *PLoS Comput. Biol.* 4(10):1–10.

[11] Seredin, O. S. Linear methods of pattern recognition for sets of objeacts of arbitrary kinds, represented by pairwise compariosions. The general case. *Proceedings of Tula State University. Natural sciences ser.* 1:141–152.

[12] Aizerman, M. A., E. M. Braverman, L. I. Rozonoer. 1970. *Metod potentsial'nykh funktsiy v teorii obucheniya mashin* [Potential functions method in machine learning theory]. Moscow: Nauka. 384 p.

[13] Sulimova, V. V. 2009. Potentsial'nye funktsii dlya analiza signalov i simvol'nykh posledovatel'nostey raznoy dliny [Kernel functions for analysis of signals and symbolic sequences of different length]. Tula. PhD Thesis. 122 p.

[14] Mottl, V. V. 2003. Metricheskie prostranstva, dopuskayushchie vvedenie lineynykh operatsiy i skalyarnogo proizvedeniya [Metric spaces admitting linear operations and inner product]. *Dokl. RAS* 388(3):312–315.

[15] Leslie, C., E. Eskin, and W. Noble. 2002. The spectrum kernel: A string kernel for SVM protein classification. *Pacific Symposium on Biocomputing Proceedings*. 564–575.

[16] Qiu, J., M. Hue, A. Ben-Hur, J.-P. Vert, and W. S. Noble. 2007. A structural alignment kernel for protein structures. *Bioinformatics* 23(9):1090–1098.

[17] Sun, L., S. Ji, and J. Ye. 2008. Adaptive diffusion kernel learning from biological networks for protein function prediction. *BMC Bioinformatics* 9(1):1–14. doi: 10.1186/1471-2105-9-162. Available at: `http://www.biomedcentral.com/1471-2105/9/162` (accessed December 27, 2016).

[18] Miklós, I., A. Novák, R. Satija, R. Lyngsø, and J. Hein. 2009. Stochastic models of sequence evolution including insertion-deletion events. *Stat. Methods Med. Res.* 18:453–485. Available at: `http://ramet.elte.hu/~miklosi/StatAlignReview2008.pdf` (accessed December 27, 2016).

[19] Onodera T., and T. Shibuya. 2013. The gapped spectrum kernel for support vector machines. *9th Conference (International) MLDM Proceedings*. Berlin – Heidelberg – New York: Springer. 1–15. doi: 10.1007/978-3-642-39712-7_1.

[20] Baisero, A., F. T. Pokorny, and C. H. Ek. 2015. On a family of decomposable kernels on sequences. *CoRR*. arXiv:/1501.06284.

[21] Seeger, M. 2009. Covariance kernels from baye ian generative models. *Stat. Methods Med. Res.* 18:453–485.

[22] Abramov, V. I., O. S. Seredin, and V. V. Sulimova. 2012. Metod opornykh ob"ektov dlya obucheniya raspoznavaniyu obrazov v evklidovykh metricheskikh prostranstvakh [Method of support objects for pattern recognition in Euclidean metric spaces]. *Conference (International) "Intellectualization of Data Processing"*. Montenegro. 5–8.

[23] Abramov, V. I., O. S. Seredin, and V. V. Mottl. 2013. Obuchenie raspoznavaniyu obrazov po metodu opornykh ob"ektov v evklidovykh metricheskikh prostranstvakh s affinnymi operatsiyami [Pattern recognition training with support vector object method in Euclidean metric spaces with

affine operations]. *Proceedings of Tula State University. Natural sciences ser.* Tula: TSU. 2(1):119–136.

[24] Hein, M., O. Bousquet, and B. Schölkopf. 2005. Maximal margin classification for metric spaces. *J. Comput. Syst. Sci.* 71(3):333–359.

[25] Xu, W. 2012. Non-Euclidean dissimilarity data in pattern recognition. Ph.D. Thesis.

[26] Seredin, O. S., V. I. Abramov, and V. V. Mottl. 2014. Affinnye operatsii v psevdoevklidovom lineynom prostranstve [Affine operations in pseudoeuclidean linear space]. *Proceedings of Tula State University. Natural sciences ser.* Tula: TSU. 3:178–196.

[27] Hancock, E. R., E. Xu, and R. C. Wilson. 2014. Pattern recognition with non-Euclidean similarities. *Man–Machine Interactions* 3:3–15.

[28] Seredin, O. S., and V. V. Mottl. 2015. Metod opornykh ob"ektov dlya obucheniya raspoznavaniyu obrazov v proizvol'nykh metricheskikh prostranstvakh [Support object method for pattern recognition training in arbitrary metric spaces]. *Proceedings of Tula State University. Natural sciences ser.* Tula: TSU. 4:178–196.

[29] Pekalska, E. M. 2005. Dissimilarity representations in pattern recognition. Concepts, theory and applications. PhD Thesis. 344 p.

[30] Bellet, A., A. Harbrad, and M. Sebban. 2013. A survey on metric learning for feature vectors and structured data. *CoRR.* abs/1306.6709. Available at: `http://arxiv.org/abs/1306.6709` (accessed December 27, 2016).

[31] Wang, J., K. Sun, F. Sha, S. Marchand-Maillet, and K. Kalousis. 2014. Two-stage metric learning. *31st Conference (International) on Machine Learning Proceedings, Cycle 2.* JMLR.org. 32:370–378. Available at: `http://jmlr.org/proceedings/papers/v32/wangc14.html` (accessed December 27, 2016).

[32] Xing, E. P., A. Y. Ng, M. I. Jordan, and S. Russel. 2003. Distance metric learning, with application to clustering with side-information. *Advances in neural information processing systems 15.* Eds. S. Becker, S. Thrun, and K. Obermayer. MIT Press. 521–528. Available at: `http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.pdf` (accessed December 27, 2016).

[33] Schultz, M., and T. Joachims. 2004. Learning a distance metric from relative comparisons. *Advances in neural information processing systems 16.* Eds. S. Thrun, L. K. Saul, and B. Schölkopf. MIT Press. 41–48. Available at: `http://papers.nips.cc/paper/2366-learning-a-distance-metric-from-relative-comparisons.pdf` (accessed December 27, 2016).

[34] Wang, J., H. Do, A. Woznica, and A. Kalousis. 2011. Metric learning with multiple kernels. *Advances in neural information processing systems 24.* Eds. J. Shawe-Taylor, R. S. Zemel, and P. L. Bartlett. MIT Press. 1170–1178. Available at: `http://papers.nips.cc/paper/4399-metric-learning-with-multiple-kernels.pdf` (accessed December 27, 2016).

[35] Dayhoff, M., R. Schwarts, and B. Orcutt. 1978. A model of evolutionary change in proteins. *Atlas of protein sequences and structures.* National Biometrical Research Foundation. 5(3):345–352.

[36] Virus Database at University College London (VIDA). Available at: `http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA3/VIDA.html` (accessed December 27, 2016).

[37] McGeoch, D. J., F. J. Rixon, and A. J. Davison. 2006. Topics in herpesvirus genomics and evolution. *Virus Res.* 117:90–104. doi: 10.1016/j.virusres.2006.01.002.

Received August 31, 2016

Исследование эффективности некоторых линейных методов классификации на модельных распределениях*

В. М. Неделько
nedelko@math.nsc.ru
Институт математики им. С. Л. Соболева СО РАН
Россия, г. Новосибирск, пр. акад. Коптюга, д. 4

Рассматривается проблема построения вероятностных моделей, позволяющих выявлять свойства методов построения решающих функций и проводить исследование этих методов. В частности, ставилась задача построения моделей, на которых заданный метод наиболее эффективен среди сравниваемых методов. Для метода логистической регрессии были построены модели, на которых этот метод эквивалентен методу максимального правдоподобия (ММП). Для метода SVM (support vector machine) построена модель, на которой этот метод приближенно эквивалентен ММП. Для дискриминанта Фишера подобной модели построить не удалось. Проведенное исследование демонстрирует перспективность подхода, основанного на построении набора «эталонных» вероятностных моделей, для исследования и сравнения методов построения решающих функций. Под эталонной моделью понимается вероятностная модель, на которой наиболее выраженно проявляется некоторое свойство исследуемого метода, например модель, на которой метод демонстрирует наибольшее превосходство, или модель, на которой проявляется некоторый недостаток метода (например, неустойчивость к «выбросам»). Также выявлены некоторые неочевидные свойства метода SVM и особенности его поведения, учет которых позволяет более эффективно применять данный метод.

Ключевые слова: *машинное обучение; логистическая регрессия; дискриминант Фишера; метод максимального правдоподобия*

DOI: 10.21469/22233792.2.3.04

1 Введение

Полезность методов построения решающих функций определяется, главным образом, точностью, которой они достигают при решении практических задач анализа данных. При этом известно, что на разных задачах лучшие результаты могут давать разные методы.

Общепринятым способом сравнения эффективности методов анализа данных является их исследование на задачах репозитория UCI. Однако такой подход обладает рядом недостатков. Первый недостаток состоит в том, что задачи попадают в репозиторий «случайно» в том смысле, что включение задачи в репозиторий зависит не от ее свойств, а от стечения обстоятельств.

Вместе с тем, предпринимались попытки создания подобного репозитория путем целенаправленного подбора задач, в частности была реализована идея включения в репозиторий для каждого из наиболее известных методов хотя бы по одной задаче, на которой этот метод был бы наиболее эффективен [1].

Другим недостатком репозиториев реальных задач является то, что каждая задача в них представлена единственной выборкой, зачастую небольшого объема, в силу чего получаемые выводы не являются строго достоверными, а носят вероятностный характер.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 14-01-00590 и № 14-07-00249.

Возникает естественная идея составить репозиторий, в котором в качестве задач были бы распределения. При этом очевидно, что придумывать распределение «наугад» неконструктивно.

В данной работе будут исследоваться возможности целенаправленного конструирования таких распределений, с тем чтобы на полученном наборе задач можно было наиболее полно выявить особенности различных методов построения решающих функций.

В качестве исследуемых методов классификации выбраны линейные методы.

В настоящее время существует несколько методов классификации, использующих линейные решающие функции [2]. Наиболее известные из них: дискриминант Фишера, машина опорных векторов (SVM) и логистическая регрессия.

Несмотря на широкое использование перечисленных методов, остается открытым вопрос, какой из методов эффективнее для тех или иных задач [3].

То, что эти методы существенно различны, дает основания предполагать, что классы задач, на которых они эффективны, также существенно различаются.

Заметим, что дискриминант Фишера и SVM в определенном смысле ближе друг к другу, чем к логистической регрессии, поскольку оба являются непараметрическими и не предполагают не только вид распределений, но и вероятностную природу данных. В отличие от этого, логистическая регрессия основана на задании определенного вида функции условной вероятности. Исходя из этого можно ожидать, что дискриминант Фишера и SVM будут эффективны на более широком классе задач [4]. Вместе с тем следует заметить, что дискриминант Фишера по факту оказывается практически идентичным решению, получаемому в предположении нормальности распределений и равенства ковариационных матриц для классов. Тот факт, что метод, полученный из сильных вероятностных предположений, указывается практически совпадающим с методом, полученным из чисто геометрических эвристик, свидетельствует о том, что класс задач, на которых метод является эффективным [5], может быть радикально шире класса задач, на которые он изначально ориентирован.

В данной работе сделана попытка подобрать семейства вероятностных моделей, которые бы позволили проявить различия в эффективности методов и проанализировать причины [6] различия эффективности.

Под вероятностной моделью в данной работе будем понимать распределение (или семейство распределений) в пространстве переменных (включая целевую), сконструированное (выбранное) в соответствии с целями проводимого исследования.

2 Основные понятия

Для введения основных понятий рассмотрим сначала общую постановку задачи построения решающих функций [7].

2.1 Постановка задачи классификации

Пусть X — пространство значений переменных, используемых для прогноза, а Y — пространство значений прогнозируемых переменных, и пусть C — множество всех вероятностных мер на мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство $\langle D, \mathfrak{B}, \mathsf{P}_c \rangle$, где $\mathfrak{B}$ — σ-алгебра, P_c — вероятностная мера. Параметр c будем называть стратегией природы.

В качестве значений целевой переменной возьмем множество $Y = \{-1, 1\}$. Как уже говорилось, мы рассматриваем случай двух классов. Для обозначения классов можно выбрать любые числовые (и нечисловые) значения. Значения -1 и 1 выбраны из соображе-

ний удобства использования в дальнейших формулах и являются одним из общепринятых вариантов.

Решающей функцией называется соответствие $\lambda\colon X \to Y$. Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$. Под риском [8] будем понимать средние потери:

$$R(c, \lambda) = \mathsf{E}_c \mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x))\ \mathsf{P}_c(dx, dy), \quad x \in X,\ y \in Y.$$

В данной работе будем рассматривать самую простую и наиболее распространенную функцию потерь $\mathcal{L}(y, y') = I(y \neq y')$. В этом случае риск есть вероятность ошибочной классификации. Здесь $I(\cdot)$ — индикаторная функция (равна 1, когда условие истинно, и 0, когда ложно).

Пусть $V_N = \{(x^i, y^i) \in D \mid i = 1, \ldots, N\}$, $V_N \in D^N$ — случайная независимая выборка из распределения P_c. В дальнейшем объем выборки N будет, как правило, фиксированным, поэтому этот параметр в обозначении выборки обычно будем опускать.

Метод (алгоритм) построения решающих функций есть отображение $Q\colon \mathcal{V} \to \Lambda$, где Λ — заданный класс решающих функций, $\mathcal{V} = \bigcup\limits_{N=1}^{\infty} D^N$ — множество всевозможных выборок, а $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q. Задача распознавания образов в стандартной постановке заключается в выборе и обосновании [9] подходящего алгоритма [10] построения решающих функций.

Критерием качества метода классификации может служить средний риск

$$\mathcal{F}_N(c, Q) = \mathsf{E}_V R(c, \lambda_{Q,V}).$$

Усреднение производится по выборкам объема N.

В случае индикаторной функции потерь качество метода характеризуется математическим ожиданием вероятности ошибочной классификации.

2.2 Оценивание функции условной вероятности

Задача распознавания образов в стандартной постановке заключается в построении и обосновании [11] подходящего алгоритма построения решающих функций. Будем также рассматривать более общую постановку задачи, когда под решающей функцией понимается оценка $\tilde{g}(x)$ функции условной вероятности

$$g(x) = \mathsf{P}_c(y = 1 \mid x) = \frac{\mathsf{P}_c(dx, y = 1)}{\mathsf{P}_c(dx)}.$$

Если в качестве значений целевой переменной выбрать 0 и 1, то функция $g(x)$ была бы функцией регрессии [12], т. е. условным математическим ожиданием (что объясняет название метода логистической регрессии). При нашем выборе значений Y функция $g(x)$ формально регрессией не является, но мы, тем не менее, следуя традиции, будем говорить о восстановлении регрессии, поскольку содержание задачи при изменении обозначений классов не меняется.

Качество решения $\tilde{g}(x)$ можно определять как вероятность ошибочной классификации для некоторого порогового классификатора, построенного по $\tilde{g}(x)$. Однако если нас интересует не просто классификация, а точность оценивания условной вероятности, то меру

качества решения $\tilde{g}(x)$ следует определять как ее точность в качестве оценки $g(x)$. Наиболее естественным представляется определение качества через использование следующей логарифмической функции потерь:

$$\mathcal{L}_g(y, \tilde{g}(x)) = -I(y = 1) \ln \tilde{g}(x) - I(y = -1) \ln(1 - \tilde{g}(x)),$$

где $I(\cdot)$ — индикаторная функция (принимает значение 1, если условие истинно, и 0 — если ложно). Выборочное среднее данной функции потерь есть взятая со знаком минус функция правдоподобия выборки по отношению к оценке условной вероятности. Математическое ожидание $R(c, \tilde{g}) = \mathsf{E}_c \mathcal{L}_g(y, \tilde{g}(x))$ этой функции потерь характеризует степень отличия $\tilde{g}(x)$ от $g(x)$, т. е. содержательно может интерпретироваться как погрешность оценки $\tilde{g}(x)$.

2.3 Эмпирические функционалы качества

Заметим, что большинство методов классификации сводятся к минимизации некоторого эмпирического критерия в заданном классе решающих функций, а многие эмпирические критерии могут быть представлены как выборочное среднее эмпирической функции потерь. Такое представление полезно для сравнения и систематизации методов построения решающих функций. При этом обоснование выбора эмпирического функционала является нетривиальной задачей [13, 14].

Простейший пример эмпирического функционала — эмпирический риск, который определяется как средние потери на выборке

$$\widetilde{R}(V, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda(x^i)).$$

Усреднение логарифмической функции потерь по выборке дает критерий правдоподобия:

$$\widetilde{R}_{MMP}(V, \tilde{g}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_g(y^i, \tilde{g}(x^i)).$$

Последнее выражение есть общепринятая функция правдоподобия, взятая с противоположным знаком. Такой знак выбран для единообразия, чтобы критерий был на минимум.

Заметим, что функция потерь $\mathcal{L}(\cdot)$ является частью постановки задачи классификации, т. е. задана, в то время как эмпирическая функция потерь, которую будем обозначать $\widetilde{\mathcal{L}}(\cdot)$, вовсе не обязана с ней совпадать и может выбираться произвольной. Поэтому эмпирический критерий качества может выглядеть как

$$\widetilde{R}(V, \tilde{g}) = \frac{1}{N} \sum_{i=1}^{N} \widetilde{\mathcal{L}}(y^i, \tilde{g}(x^i)).$$

Заметим, что в таком виде эмпирический критерий представляется не всегда (контрпримером является дискриминант Фишера). Кроме того, критерий может содержать регуляризатор.

Функцию $\widetilde{\mathcal{L}}(\cdot)$ часто называют функцией потерь. Однако мы функцией потерь называем величину $\mathcal{L}(y, \lambda(x))$ — это функция, которая является неотъемлемой частью постановки задачи классификации и отражает объективные потери от неверного решения. Функция потерь задается «заказчиком», т. е. является внешними требованиями.

В отличие от нее функция $\widetilde{\mathcal{L}}(\cdot)$ является эвристикой и частью метода решения задачи. Эта функция может выбираться произвольно, на основе интуиции исследователя. Будем называть эту функцию эмпирической функцией потерь.

Связь между этими функциями заключается в том, что разработчик метода ожидает, что минимизация эмпирической функции потерь на выборке в определенной степени соответствует минимизации функции потерь на (неизвестном) распределении (контрольных выборках). При этом ожидание, как правило, основывается на интуиции либо на полуформальной аргументации.

3 Линейные методы классификации

Пусть X представлено декартовым произведением количественных переменных $X = \prod_{j=1}^{n} X_j$, тогда $x \in X$ представляет собой n-мерный вектор.

3.1 Логистическая регрессия

Метод логистической регрессии [15] подразумевает параметрическое оценивание (построение оценки $\tilde{g}(x)$) функции условной вероятности $g(x)$.

Для вывода модели логистической регрессии рассмотрим случай нормальных распределений классов с равными ковариационными матрицами S, т. е. условные меры $\mathsf{P}_c(dx \mid y)$ задаются плотностями:

$$\varphi_y(x) = \frac{1}{(2\pi)^{n/2}|S|^{1/2}} e^{-0{,}5(x-\mu_y)^T S^{-1}(x-\mu_y)}.$$

Имеем

$$g(x) = \mathsf{P}_c(y = 1 \mid x) = \frac{p\varphi_1(x)}{p\varphi_1(x) + (1-p)\varphi_{-1}(x)},$$

где $p = \mathsf{P}_c(y = 1)$ — безусловная вероятность первого класса.

Подставив нормальную плотность, после элементарных преобразований получаем:

$$g(x) = \frac{1}{1 + e^{-(w'x + w'_0)}} = \sigma(w'x + w'_0),$$

где $w' = 0{,}5S^{-1}(\mu_1 - \mu_{-1})$, $w'_0 = 0{,}5\mu_{-1}^T S^{-1}\mu_{-1} - 0{,}5\mu_1^T S^{-1}\mu_1 + \ln p - \ln(1-p)$.

Здесь $\sigma(z) = 1/(1 + e^{-z})$ — так называемая логистическая функция (иногда также называемая сигмоидом или логит-функцией). Логистическая функция широко используется в методах анализа данных, в частности как функция активации в нейронных сетях, а также как функция для замены переменных при необходимости перейти от ограниченного множества значений (например, от вероятностей) к неограниченному множеству значений (например, для построения линейной регрессии) и наоборот.

Метод логистической регрессии основан на оценивании функции условной вероятности моделью $\tilde{g}(x) = \sigma(wx + w_0)$, в которой w и w_0 — настраиваемые параметры. Знак транспонирования для вектора в записи скалярного произведения будем опускать.

Как следует из выкладок, модель логистической регрессии является точной для случая нормальных распределений с равными ковариационными матрицами. Очевидно, что она является точной и для гораздо более широкого класса распределений, поскольку она определяет только условное распределение, но никак не зависит от безусловного распределения в X.

На практике параметры модели обычно оцениваются путем максимизации критерия правдоподобия

$$\mathfrak{R}_\sigma(V,\tilde{g}) = \frac{1}{N}\sum_{i=1}^{N} -I(y^i=1)\ln\tilde{g}(x^i) - I(y^i=-1)\ln(1-\tilde{g}(x^i)).$$

Учитывая, что $1-\sigma(z)=\sigma(-z)$, предыдущее выражение можно привести к виду:

$$\mathfrak{R}_\sigma(V,w,w_0) = \frac{1}{N}\sum_{i=1}^{N} -\ln\sigma(-y^i(wx^i+w_0)) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathcal{L}}(y^i(wx^i+w_0)).$$

Здесь $\widetilde{\mathcal{L}}(z) = -\ln\sigma(-z)$ — эмпирическая функция потерь.

3.2 Метод опорных векторов

Метод опорных векторов ранее был известен как метод обобщенного портрета.

Так же как и в дискримнанте Фишера, идея метода опорных векторов заключается в поиске такого направления в пространстве переменных, по которому классы были бы наиболее разделимы. Отличие заключается в критерии качества разделимости.

В методе опорных векторов мера разделимости основывается на понятии зазора, который понимается как ширина разделяющей полосы между классами.

Запишем линейный пороговый классификатор в виде:

$$\lambda(x) = \operatorname{sign}(wx - w_0).$$

Требуется найти вектор w и скаляр w_0, минимизирующие эмпирический риск и одновременно максимизирующие ширину разделяющей полосы.

Если классы линейно разделимы, то задача может быть записана как задача максимизации зазора при ограничениях, обеспечивающих безошибочную классификацию обучающей выборки.

Поскольку норма вектора w не влияет на направление, выберем удобную нормировку из условия

$$\min_{(x^i,y^i)\in V} y^i(x^iw - w_0) = 1.$$

Для граничных точек имеем:

$$x_+w - w_0 = 1; \quad -(x_-w - w_0) = 1.$$

Ширина разделяющей полосы:

$$(x_+ - x_-)\frac{w}{|w|} = \frac{(w_0+1)-(w_0-1)}{|w|} = \frac{2}{|w|}.$$

Получаем задачу квадратичной оптимизации:

$$\begin{cases} w^2 \to \min\limits_{w,w_0}; \\ y^i(x^iw - w_0) \geqslant 1\,, \quad i=1,\ldots,N\,. \end{cases}$$

Условие нормировки выполняется автоматически.

В случае линейно неразделимой выборки метод также сводится к задаче квадратичного программирования, которая выглядит как

$$\begin{cases} \dfrac{w^2}{2} + C\sum\limits_{i=1}^{N}\xi_i \to \min\limits_{w,w_0,\xi}; \\ y^i(x^i w - w_0) \geqslant 1 - \xi_i; \\ \xi_i \geqslant 0, \quad i = 1,\ldots,N, \end{cases}$$

где $C > 0$ — параметр.

Задача эквивалентна задаче безусловной минимизации (по w и w_0) следующего эмпирического критерия:

$$\mathfrak{R}(V, w, w_0) = \frac{w^2}{2} + C\sum_{i=1}^{N}\big(1 - y^i(x^i w - w_0)\big)_+ = \frac{w^2}{2} + C\sum_{i=1}^{N}\widetilde{\mathcal{L}}(y^i(wx^i + w_0)).$$

Здесь $(z)_+ = zI(z > 0)$ обозначает функцию, которая «зануляет» отрицательные значения аргумента.

Критерий приведен к виду с эмпирической функцией потерь $\widetilde{\mathcal{L}}(z) = (1 - z)_+$.

Слагаемое $w^2/2$ имеет смысл регуляризатора (т. е. слагаемого, вводимого для повышения устойчивости решений).

Такая форма выявляет большое сходство SVM и логистической регресии [16], которое становится особенно наглядным, если сравнить графики эмпирических функций потерь для этих методов (рис. 1).

Это сходство представляется довольно неожиданным ввиду того, что SVM — непараметрический метод, а логистическая регрессия — параметрический (с неполной вероятностной моделью).

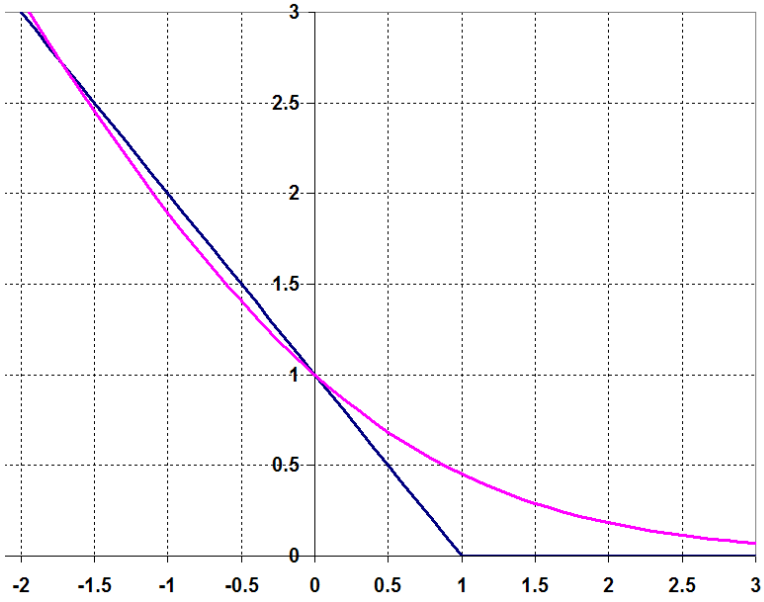


Рис. 1 Эмпирические функции потерь для SVM и логистической регрессии

Нам представляется более удобным записать критерий в несколько ином виде:

$$\mathfrak{R}(V, w, w_0) = \varkappa \frac{w^2}{2} + \frac{1}{N} \sum_{i=1}^{N} \widetilde{\mathcal{L}}(y^i(wx^i + w_0)).$$

Здесь вместо параметра C введен параметр $\varkappa = N/C$. Преимущество этой формы записи критерия состоит в том, что нулевое значение параметра C не имеет смысла, в то время как $\varkappa = 0$ – вполне допустимое и разумное значение параметра, которое соответствует отсутствию регуляризатора. При этом следует оговориться, что для случая линейно разделимой выборки $\varkappa = 0$ все же не годится, поскольку решение будет определяться неоднозначно (и не будет максимизироваться зазор), однако остаются допустимыми сколь угодно малые значени $\varkappa$.

Утверждение 1. *Оптимальное значение критерия SVM не превосходит* 1, *т. е.*

$$\min_{w, w_0, \xi} \mathfrak{R}(V, w, w_0) \leqslant 1.$$

Доказательство. Рассмотрим решение с параметрами

$$w = 0, \quad w_0 = \arg\max_{y} \sum_{i=1}^{N} I(y^i = y),$$

т. е. w_0 принимает значение того класса, представителей которого в выборке больше (при равенстве частот — любого). Эта решающая функция для всех x прогнозирует класс с наибольшей частотой в обучающей выборке.

Очевидно, что эмпирический риск для такой решающей функции не превосходит 0,5. При этом непосредственной подстановкой можно убедиться, что значение критерия SVM для такого решения есть удвоенное значение эмпирического риска. ■

Данная простая оценка объясняет, почему в случае, когда классы плохо разделимы, метод SVM предпочитает их не разделять вовсе, а относить все объекты к одному классу. Особенно часто это происходит, когда частоты классов в обучающей выборке сильно различаются.

Заметим, что для подобных решающих функций, которые все объекты относят к одному классу, ширина раздеряющей полосы (зазор) формально бесконечна. Увеличение параметра $\varkappa$ повышает вероятность того, что метод SVM построит такое решение.

3.3 Дискриминант Фишера

Дискриминант Фишера, как и SVM, использует исключительно метрические свойства конфигурации выборочных точек и не требует не только никаких предположений о распределениях, но и вообще статистической постановки задачи классификации.

Идея дискриминанта Фишера заключается в выборе такого направления в пространстве переменных, при проецировании выборки на которое образы классов оказываются в некотором смысле наиболее удаленными друг от друга. Формально это выражается в максимизации следующего критерия:

$$\Phi(w) = \frac{(\tilde{\mu}_{w,1} - \tilde{\mu}_{w,-1})^2}{\tilde{S}_w},$$

где $\tilde{\mu}_{w,y} = (1/N_y) \sum_{i=1}^{N} wx^i I(y^i = y)$ — среднее проекций точек выборки y-го класса на направление w, N_y — число объектов y-го класса в выборке, а $\tilde{S}_w = \sum_{i=1}^{N} (wx^i - \tilde{\mu}_{w,y^i})^2$ —

суммарный квадрат отклонений проекций точек выборки y-го класса на направление w от среднего этого класса,

Критерий приводится к виду:

$$\Phi(w) = \frac{(w\tilde{\mu}_1 - w\tilde{\mu}_{-1})^2}{w^T\tilde{S}w} = \frac{w^T(\tilde{\mu}_1 - \tilde{\mu}_{-1})(\tilde{\mu}_1 - \tilde{\mu}_{-1})^T w}{w^T\tilde{S}w},$$

где $\tilde{\mu}_{w,y} = (1/N_y)\sum_{i=1}^{N} x^i I(y^i = y)$ — среднее точек выборки y-го класса, $\tilde{S}_y$ — выборочная ковариационная матрица y-го класса, $\tilde{S} = N_1\tilde{S}_1 + N_{-1}\tilde{S}_{-1}$.

Последняя форма критерия имеет вид отношения Релея.

Известно, что максимум $\Phi(w)$ достигается при $w = w_\Phi = \tilde{S}^{-1}(\tilde{\mu}_1 - \tilde{\mu}_{-1})$.

Заметим, что выражение для w_Φ совпадает (с точностью до нормировки) с выражением для нормали к разделяющей гиперплоскости для случая нормальных распределений с равными матрицами ковариаций. Такое сходство приводит к тому, что в литературе эти методы иногда смешиваются, несмотря на их принципиальное различие по подходу и предположениям.

После выбора направления w_Φ задача классификации становится одномерной и может быть достаточно легко решена. Более того, в некоторых случаях полученное (проецированием) упорядочивание объектов само по себе может считаться решением.

3.4 Модификации методов

Рассмотренные методы обладают схожими чертами, в частности они допускают одинаковые усовершенствования.

Первое усовершенствование касается регуляризации.

Заметим, что SVM регуляризующий член содержит изначально — это $\varkappa w^2/2$, слагаемое, отвечающее за максимизацию отступа (зазора).

Ровно это же слагаемое можно добавить к критерию логистической регрессии. И это слагаемое также будет отвечать за максимизацию отступа. Хотя отступ для логистической регрессии имеет не такой очевидный смысл, как для SVM, его можно определить как ширину полосы между классами, в которой эмпирическая функция потерь для обоих классов не превышает 1. Это определение годится и для SVM.

Для дискриминанта Фишера аналогичный регуляризатор выглядит как добавка к матрице S в виде некоторой диагональной матрицы (аналогично гребневой регрессии).

Второе усовершенствование касается использования функции ядра (так называемый kernel trick). Заметим, что исторически kernel trick был изначально разработан для метода SVM и лишь впоследствии распространен на остальные методы. Тем не менее, этот прием [17] полностью аналогичен для всех трех методов.

Возможность перехода к ядрам обусловлена двумя факторами. Во-первых, линейностью решения, т. е. тем, что решение дается через скалярное произведение вектора, описывающего объект, с направляющим вектором. Во-вторых, критерии всех трех методов таковы, что оптимальное направление может быть представлено как линейная комбинация векторов выборки. Это справедливо не только в том очевидном случае, когда ранг системы векторов выборки равен размерности пространства, но и в общем случае, в том числе для бесконечномерных пространств.

Из этих фактов элементарно следует, что решение может быть записано с использованием только скалярных произведений между объектами, но не требует даже признаковых описаний объектов [18].

Если бы второй факт не имел места, то оптимальное направление в спрямляющем пространстве могло бы не иметь прообраза в исходном пространстве, и тогда это спрямляющее пространство пришлось бы вводить явно.

4 Решающие функции на распределениях

Естественным первым шагом в оценивании эффективности метода построения решающих функций является исследование метода на распределениях. Фактически это означает изучение асимптотических свойств [19, 20] метода (при объеме выборки, стремящемся к бесконечности).

Применение методов к распределениям позволяет, например, оценить, насколько в принципе метод способен приблизиться к Байесовскому решению. В данной работе это также позволит аналитически исследовать поведение отступа.

В отличие от работ, в которых вероятностная модель конструируется на основе выборки [21], будем конструировать модели таким образом, чтобы исследуемые особенности методов проявлялись на них наиболее выраженно.

В качестве первого примера рассмотрим модель нормальных распределений с равными ковариационными матрицами. Положим для простоты безусловные вероятности классов равными, т. е. $\mathsf{P}(y) = 0{,}5$.

Для удобства выберем начало координат посредине между центрами классов, т. е. так, чтобы для векторов математических ожиданий выполнялось $\mu_{-1} = -\mu_1$. Введем параметр μ, через который выразим вектора математических ожиданий как $\mu_y = y\mu$.

Вычислим для данной модели значение критерия SVM:

$$\mathfrak{R}(c, w, w_0) = \varkappa\frac{w^2}{2} + \int\limits_D \widetilde{\mathcal{L}}(y(wx + w_0))\ \mathsf{P}_c(dx, dy).$$

При сделанном выборе начала координат оптимальное значение параметра w_0 равно 0, поэтому w_0 можно исключить из выражения.

Рассмотрим для начала одномерный случай, когда x, w и μ — скаляры.

Для модели с одномерными нормальными распределениями можем записать:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \sum_{y\in\{-1,1\}} \frac{\mathsf{P}(y)}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-(x-\mu_y)^2/(2\sigma)}(1 - y(wx))_+\, dx.$$

Учитывая симметрию распределений классов, после преобразований получаем:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{1}{w}} e^{-(x-\mu)^2/(2\sigma)}(1 - wx)\, dx,$$

Делая стандартную замену $t = (x - \mu)/\sigma$, получаем:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{(1-\mu w)/(w\sigma)} e^{-t^2/2}(1 - w(t\sigma + \mu))\, dt.$$

После элементарных преобразований выражение приводится к виду:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \frac{zF_{\mathcal{N}}(z) + F'_{\mathcal{N}}(z)}{z + m},$$

где $m = \mu/\sigma$; $z = 1/(w\sigma) - m$; $F_{\mathcal{N}}(\cdot)$ — функция (интегральная) стандартного нормального распределения; $F'_{\mathcal{N}}(\cdot)$ — плотность нормального распределения.

Полученное выражение позволяет установить некоторые свойства SVM.

Введем обозначение:

$$w^* = \arg\min_w \mathfrak{R}(c, w).$$

Величина $1/w^*$ называется зазором или отступом (margin).

Введем обозначение $s = 1/(w^*\sigma)$.

Утверждение 2. *Величина зазора s для метода SVM без регуляризации (при $\varkappa = 0$) монотонно убывает с ростом m.*

Данное свойство выглядит парадоксальным: при отдалении распределений классов друг от друга величина зазора (т. е. ширина разделяющей полосы) уменьшается.

Доказательство. Пусть $\varkappa = 0$. Для нахождения w^* вычислим производную:

$$\frac{\partial \mathfrak{R}(c, w)}{\partial z} = \frac{mF_{\mathcal{N}}(z) - F'_{\mathcal{N}}(z)}{(z+m)^2}.$$

Производная равна нулю при $m = F'_{\mathcal{N}}(z^*)/F_{\mathcal{N}}(z^*)$, где $z^* = s - m$.

Требуется установить, что зависимость s от m, неявно задаваемая выражением $m = F'_{\mathcal{N}}(s-m)/F_{\mathcal{N}}(s-m)$, является монотонно убывающей функцией. Для этого вычислим производную $m' = dm/ds$. Получаем:

$$m' = \frac{F''_{\mathcal{N}}(s-m)F_{\mathcal{N}}(s-m) - (F'_{\mathcal{N}}(s-m))^2}{(F_{\mathcal{N}}(s-m))^2}(1-m') = -sm(1-m'),$$

откуда выражаем

$$m' = \frac{sm}{sm-1}.$$

Легко убедиться, что $sm < 1$. Действительно,

$$sm = (z+m)m = \left(z + \frac{F'_{\mathcal{N}}(z)}{F_{\mathcal{N}}(z)}\right)\frac{F'_{\mathcal{N}}(z)}{F_{\mathcal{N}}(z)}.$$

Остается убедиться, что

$$zF'_{\mathcal{N}}(z)F_{\mathcal{N}}(z) + (F'_{\mathcal{N}}(z))^2 - (F_{\mathcal{N}}(z))^2 < 0$$

при любых z. Это можно сделать стандартным методом анализа функций, т. е. вычислив и проанализировав несколько производных.

Таким образом, установлено, что $m' < 0$. ■

Оптимальное значение критерия SVM при $\varkappa = 0$ есть

$$\mathfrak{R}(c, w^*) = \frac{z^*F_{\mathcal{N}}(z^*) + F'_{\mathcal{N}}(z^*)}{z^*+m} = F_{\mathcal{N}}(z^*).$$

Только что было доказано, что величина s убывает с ростом m, но тогда $z^* = s - m$ — тем более монотонно убывающая функция m. Учитывая монотонность $F_{\mathcal{N}}(\cdot)$, заключаем, что $\mathfrak{R}(c, w^*)$ монотонно убывает с ростом m.

Данный факт вполне естественен и очевиден. Его можно доказать намного проще, но в приведенных выкладках были также получены полезные выражения для отступа.

Вернемся теперь к случаю многомерного пространства переменных.

Проекция нормального распределения на произвольное направление w есть одномерное нормальное распределение, параметры которого обозначим μ_w и σ_w.

Значение критерия SVM по направлению w полностью определяется (при $\varkappa = 0$) величиной $m_w = \mu_w/\sigma_w$, причем монотонно от нее зависит.

Заметим, что критерий дискриминанта Фишера для рассматриваемой модели нормальных распределений есть просто $4m_w^2$.

Из сказанного следует, что оптимальные направления для разделяющих функций SVM и дискриминанта Фишера совпадают.

Утверждение 3. *При $\varkappa = 0$ на модели нормальных распределений с равными матрицами ковариаций решения методами SVM, логистической регрессии и дискриминанта Фишера совпадают с Байесовским решением.*

Данный факт является известным.

Применительно к дискриминанту Фишера это классический результат. Для логистической регрессии утверждение доказывается элементарно.

Что касается SVM, то автору не удалось найти работы, где этот метод применялся бы не к выборке, а к распределениям. Однако известны результаты о состоятельности метода SVM [22], из которых следуют схожие выводы.

Справедливость утверждения для метода SVM следует из установленного выше факта совпадения получаемого решения с решением дискриминанта Фишера.

Заметим, что исследуемые методы дают оптимальные решения далеко не на всех вероятностных моделях.

Утверждение 4. *Существуют вероятностные модели, для которых Байесовская разделяющая функция линейна, но решения, полученные методами SVM, логистической регрессии и дискриминанта Фишера, не являются оптимальными.*

Доказательство. Для доказательства явно построим модель, обладающую требуемым свойством.

Эта модель является смесью исходной модели нормальных распределений с равными матрицами ковариаций и модели с распределениями, сосредоточенными в некоторой точке $\check{x}$, причем для этой второй компоненты вероятности классов в точке $\check{x}$ одинаковы.

Очевидно, что добавление описанной компоненты к исходной модели не меняет Байесовского решения, поскольку для этой компоненты классы неразделимы (любое решение дает одинаковую вероятность ошибочной классификации).

При этом вес второй компоненты и положение точки $\check{x}$ существенно влияют на решения, получаемые перечисленными методами. В частности, метод SVM при достаточно большом весе второй компоненты изменит решение так, чтобы точка $\check{x}$ попала внутрь разделяющей полосы. ■

Выясним теперь, как влияет на решение параметр регуляризации $\varkappa$ в методе SVM.

На рис. 2 приведены разделяющие функции, построенные методом SVM для модели с нормальными распределениями. Ковариационные матрицы для обоих классов одинаковы и соответствуют стандартным отклонениям, равным по главным осям 1 и 2. Математические ожидания для классов равны соответственно -1 и 1. Распределения на рисунке изображены соответственно синим и зеленым эллипсами (кривые равной плотности вероятности). Черная прямая соответствует решению, построенному методом SVM при $\varkappa = 0$,

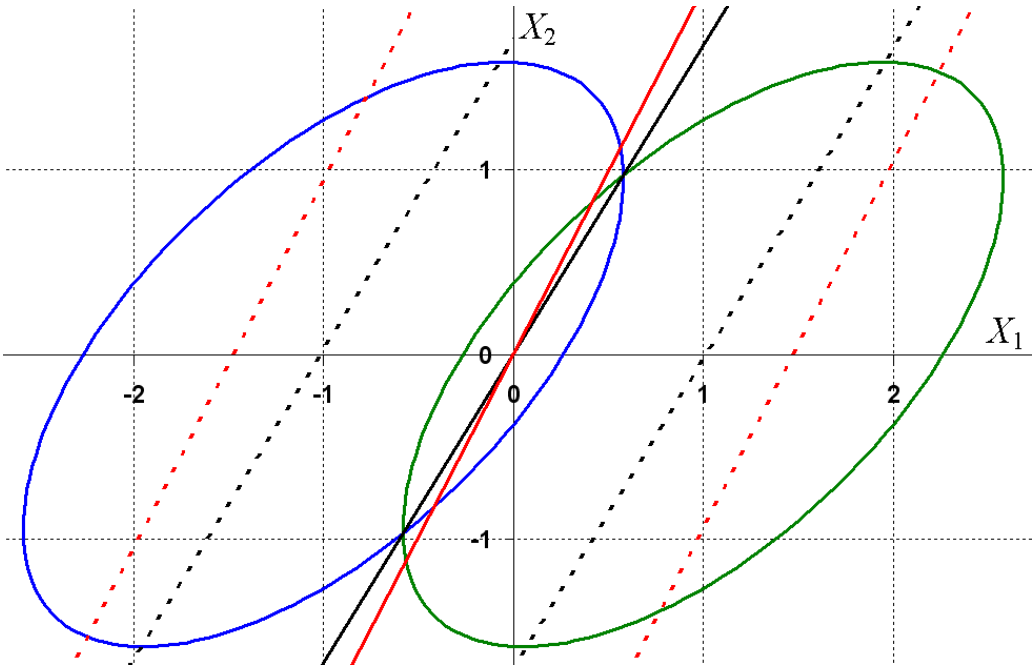


Рис. 2 Решения, построенные методом SVM для модели с нормальными распределениями при нулевом (черная линия) и ненулевом (красная линия) значениях параметра регуляризации $\varkappa$.

которое совпадает с Байесовской решающей функцией. Красная прямая соответствует $\varkappa = 0{,}2$. Пунктирные линии отмечают разделяющую полосу для решений соответствующего цвета.

Как видно из примера, увеличение $\varkappa$ приводит к повороту разделяющей прямой, как если бы эллипсы распределений стали менее вытянутыми. Такое же изменение решения можно получить и на основе дискриминанта Фишера, добавляя к оценке ковариационной матрицы единичную матрицу с некоторым коэффициентом.

5 Модели максимального правдоподобия

Для исследования эффективности методов выберем подходящие задачи, т. е. вероятностные модели, в соответствии с которыми будут генерироваться выборки для тестирования методов.

Вообще говоря, модель для тестирования можно придумать совершенно произвольную. Выберем для каждого метода модель, которая ему в некотором смысле наиболее подходит, а именно: модель, при которой метод классификации соответствует ММП (или близок к нему).

Следует уточнить, в каком смысле мы говорим о соответствии.

5.1 Метод максимального правдоподобия

В классическом виде ММП каждой выборке сопоставляет распределение (из заданного параметрического семейства), для которого функция правдоподобия данной выборки максимальна.

В более универсальном виде ММП можно сформулировать как метод, сопоставляющий выборке распределение, которое критерием отношения правдоподобия будет выбрано против любой альтернативы.

Пусть Θ — заданное множество вероятностных мер (в параметрическом случае будем отождествлять Θ со множеством значений параметров).

Будем говорить, что вероятностная мера θ_1 не менее правдоподобна, чем мера θ_2, по отношению к выборке V, если во всех точках выборки существует производная Радона–Никодима меры θ_2 по мере θ_1 и произведение этих производных по всем точкам выборки не превосходит 1.

Метод максимального правдоподобия сопоставляет выборке меру θ^*, которая не менее правдоподобна, чем любая мера из Θ.

Заметим, что ММП определен не для любого семейства распределений Θ, поскольку не в любом Θ найдётся θ^* с требуемыми свойствами. Однако классический вариант ММП, основанный на функции правдоподобия, является частным случаем приведенного.

Такое обобщение нужно, чтобы иметь возможность определить ММП для случаев, когда семейство распределений очень широкое. В частности, так можно определить ММП даже для случая всех вероятностных мер (заданных на одной и той же σ-алгебре). При этом наиболее правдоподобная мера будет эмпирическим распределением для выборки (когда в каждой точке выборки сосредоточена вероятность $1/N$).

Будем считать, что ММП задает отображение множества выборок в некоторое множество Θ, элементами которого являются в зависимости от постановки задачи либо вероятностные меры, либо значения параметров распределений. Будем обозначать это отображение как $\theta^*(V)$, где $V \in D^N$.

Определение 1. *Будем говорить, что метод классификации Q, отображающий множество выборок D^N во множество решающих функций Λ, соответствует ММП для семейства распределений Θ, если существует отображение $\zeta : \Theta \to \Lambda$, такое что $\lambda_{Q,V} = \zeta(\theta^*(V))$ для всех $V \in D^N$.*

Такое отображение, очевидно, существует тогда и только тогда, когда не существует двух выборок, образы которых для ММП совпадают, а для метода Q различаются.

Заметим, что если в качестве Θ взять множество всех вероятностных мер, то любой метод классификации будет соответствовать ММП. Поэтому факт такого соответствия практически значим только в случае, если класс Θ достаточно узок.

Определение 2. *Будем говорить, что метод классификации Q эквивалентен ММП для семейства распределений Θ, если существует взаимно однозначное отображение $\zeta : \Theta \to \Lambda$, такое что $\lambda_{Q,V} = \zeta(\theta^*(V))$ для всех $V \in D^N$.*

Попытаемся построить для рассматриваемых в работе линейных методов классификации модели, для которых эти методы будут эквивалентны методам максимального правдоподобия.

5.2 Общий вид модели

Логистическая регрессия была построена как модель максимального правдоподобия. Возникает вопрос, можно ли остальные методы интерпретировать как методы максимального правдоподобия для некоторой вероятностной модели.

Если критерий выражается через эмпирическую функцию потерь, то соответствующей вероятностной моделью может быть параметрическое семейство распределений с плотностью вида

$$\varphi(x, y) = A(w, w_0)e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}\varphi_0(x). \tag{1}$$

Идея построения этой модели очевидна. Если мы хотим, чтобы функция потерь совпадала с функцией правдоподобия (с обратным знаком), то напрашивается задать плотность вероятности как экспоненту от функции потерь. Однако сама по себе функция $e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}$

не может быть плотностью, поскольку интеграл от нее бесконечен. По этой причине приходится включить множитель $\varphi_0(x)$ — это некоторая функция, не зависящая от параметров w и w_0, которая обеспечивает конечность интеграла. Кроме того, требуется добавить нормировочный множитель $A(w, w_0)$.

В общем случае нормировочный множитель зависит от параметров w и w_0, поэтому добиться точного совпадения критерия на основе эмпирической функции потерь и критерия максимального правдоподобия не всегда удается. Однако отличие (связанное с наличием этого нормировочного множителя) на практике несущественно.

Утверждение 5. *Для того чтобы эмпирическая функция потерь была функцией правдоподобия для условной вероятности, т. е. чтобы было возможно представление* $\mathsf{P}(y \mid x) = e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}$, *необходимо и достаточно, чтобы выполнялось соотношение*

$$\widetilde{\mathcal{L}}(z) = -\ln\left(1 - e^{-\widetilde{\mathcal{L}}(-z)}\right).$$

Доказательство. Для условной вероятности должно выполняться соотношение

$$\mathsf{P}(y = 1 \mid x) = 1 - \mathsf{P}(y = -1 \mid x).$$

Подставив требуемое представление, имеем

$$e^{-\widetilde{\mathcal{L}}(+1(wx+w_0))} = 1 - e^{-\widetilde{\mathcal{L}}(-1(wx+w_0))}.$$

Элементарными преобразованиями получаем искомое. ■

Следствие 1. *При выполнении условия из утверждения* 1 *модель* (1) *может быть представлена в виде:*

$$\varphi(x, y) = e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}\varphi_0(x),$$

где $\varphi_0(x)$ *— безусловная плотность для* x.

Нормировочный множитель $A(w, w_0)$ в этом случае не требуется, поскольку, умножив плотность на условную вероятность, получим автоматически нормированную совместную плотность.

5.3 Логистическая регрессия

Логистическая регрессия изначально построена как модель максимального правдоподобия для условной вероятности. Это означает, что условие утверждения 1 должно выполняться (в чем легко убедиться непосредственно).

В качестве модели, на которой логистическая регрессия будет построена ММП, возьмем распределение

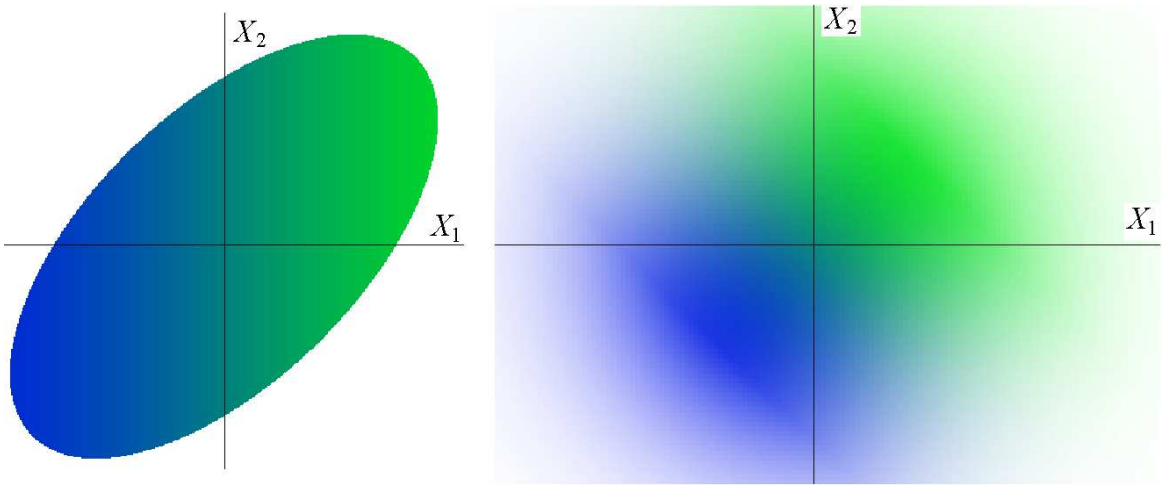
$$\varphi(x, y) = \sigma(y(wx))\varphi_{\mathcal{E}}(x),$$

где $\varphi_{\mathcal{E}}(x)$ — равномерное распределение внутри эллипса с осями 2 и 1, главная ось повернута на угол $\pi/4$.

Пример распределения из этого семейства приведен на рис. 3, *а*. Интенсивность синего и зеленого цветов отражает плотности вероятности классов 1 и -1 при $w = (1, 0)$.

5.4 Метод опорных векторов

Для метода SVM невозможно подобрать модель максимального правдоподобия [23], где бы от параметров зависела только условная вероятность.



$(а)$ $(б)$

Рис. 3 Плотности вероятности двух классов для моделей логистической регрессии $(а)$ и SVM $(б)$

Выберем следующее семейство распределений (является модификацией модели из [24]):

$$\varphi(x,y) = A\left(|w|\right) e^{-(1-y(wx))_+} \varphi_{\mathcal{N}}(x)\,, \tag{2}$$

где $\varphi_{\mathcal{N}}(x)$ — нормальное распределение с нулевым средним и единичной ковариационной матрицей.

Пример распределения из этого семейства приведен на рис. 3, *б*. Интенсивность синего и зеленого цветов отражает плотности вероятностей классов 1 и -1 при $w\sqrt{2} = (1,1)$.

Данную плотность можно представить как произведение условной вероятности

$$\mathsf{P}(y \mid x) = \sigma\left(y\left((1+wx)_+ - (1-wx)_+\right)\right)$$

и безусловной плотности

$$\varphi(x) = A\left(|w|\right)\left(e^{-(1+wx)_+} + e^{-(1-wx)_+}\right)\varphi_{\mathcal{N}}(x).$$

Для наглядности преобразуем функцию условной вероятности:

$$g(x) = \sigma\left((1+wx)_+ - (1-wx)_+\right) = \begin{cases} \sigma(wx-1), & wx < -1\,; \\ \sigma(2wx), & -1 \leqslant wx \leqslant 1\,; \\ \sigma(wx+1), & wx > 1. \end{cases}$$

Данная функция изображена синей кривой на рис. 4. Красная кривая изображает функцию $e^{-(1+wx)_+} + e^{-(1-wx)_+}$.

Вычислим $A(|w|)$. Из условия нормировки имеем:

$$\frac{1}{A\left(|w|\right)} = \sum_{y\in\{-1,1\}} \int\limits_X \varphi(x,y)\,dx = \sum_{y\in\{-1,1\}} (2\pi)^{-n/2} \int\limits_X e^{-x^2/2} e^{-(1-y(wx))_+}\,dx\,.$$

Повернем систему координат так, чтобы переменная X_1 была в направлении вектора w. Учитывая симметричность распределений при разных y, имеем:

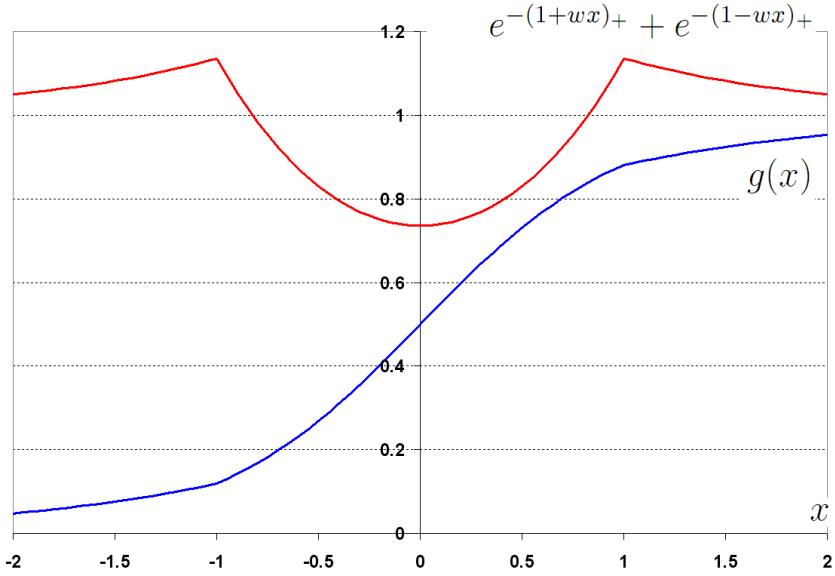


Рис. 4 Функция условной вероятности $g(x) = \sigma\left((1+wx)_+ - (1-wx)_+\right)$ и параметрическая компонента распределения $e^{-(1+wx)_+} + e^{-(1-wx)_+}$ для модели SVM

$$\frac{1}{2A(|w|)} = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-x^2/2} e^{-(1-|w|x_1))_+}\, dx_1 = e^{w^2/2-1} F_{\mathcal{N}}\left(\frac{1}{|w|} - |w|\right),$$

где $F_{\mathcal{N}}(\cdot)$ — функция (интегральная) нормального распределения.

Заметим, что при малых $|w|$ имеет место $-\ln A(|w|) \approx w^2/2 - 1 + \ln 2$. В этом случае функция правдоподобия (с обратным знаком) с точностью до аддитивной константы может быть представлена как

$$-\ln \varphi(x, y) \approx \frac{w^2}{2} + (1 - y(wx))_+.$$

Получаем, что при малых $|w|$ для рассмотренной модели критерий максимального правдоподобия совпадает с критерием SVM при $\varkappa = N$.

Заметим, что во многих источниках регуляризующее слагаемое $\varkappa w^2/2$ интерпретируется в рамках Байесовского подхода [25], когда параметр w полагается случайным [26, 27]. Здесь, однако, получена модель максимального правдоподобия для критерия SVM без вероятностной интерпретации параметра w.

5.5 Дискриминант Фишера

Труднее всего подобрать модель максимального правдоподобия для дискриминанта Фишера.

Это утверждение может показаться неожиданным, поскольку дискриминант Фишера принято связывать с вероятностной моделью нормальных распределений с равными ковариационными матрицами.

Следует, однако, уточнить, в чем заключается эта связь.

Известно, что дискриминант Фишера дает оптимальное (Байесовское) решение, будучи примененным непосредственно к нормальным распределениям с равными ковариационными матрицами. Однако, как мы выяснили в разд. 4, ровно этим же свойством обладают и логистическая регрессия, и SVM при $\varkappa = 0$.

Также связь заключается в том, что дискриминант Фишера выражается через оценки максимального правдоподобия нормальных распределений. Это означает, что дискриминант Фишера соответствует ММП на классе нормальных распределений (с равными ковариационными матрицами), но не эквивалентен ему, поскольку по решающей функции параметры вероятностной модели не восстанавливаются.

При этом критерий дискриминанта Фишера не может являться функцией правдоподобия, хотя бы потому, что он не аддитивен по объектам выборки.

Гипотеза 1. Не существует невырожденной вероятностной модели, на которой дискриминант Фишера был бы эквивалентен ММП.

Под невырожденностью здесь понимается, что размерность пространства переменных больше 1 и что вероятности не сосредоточены на многообразиях нулевой меры Лебега.

6 Численный эксперимент

Сравнение методов проводилось неоднократно (см., например, [28]), однако доступные в литературе выводы об области применимости каждого метода носят частный характер.

В ближайшее время вряд ли следует ожидать появления исчерпывающего описания семейств вероятностных моделей, для которых был бы наиболее предпочтителен заданный метод, например SVM. Задача построения такого описания представляется чрезвычайно сложной.

В данной работе численный эксперимент проводится также на конкретных частных примерах, однако вероятностные модели подбираются как характерные представители определенных классов моделей с заданными свойствами (например, модели с редкими большими уклонениями). Это позволяет надеяться, что обнаруженные закономерности в поведении исследуемых методов будут иметь более общий характер.

В качестве задач, на которых будут тестироваться методы, выбраны следующие модели.

В первую очередь, для каждого метода была сконструирована вероятностная модель, на которой этот метод предположительно должен давать наилучший результат. Для дискриминанта Фишера это модель с нормальными распределениями, для SVM и логистической регрессии это модели максимального правдоподобия.

Предварительный численный эксперимент, однако, показал, что на всех трех моделях лучшим оказывается дискриминант Фишера. В связи с этим были целенаправлено подобраны модели, позволяющие каждому методу продемонстрировать преимущество.

Список моделей:

1) нормальные распределения с равными матрицами ковариаций. Модель описана в разд. 4, параметры те же: стандартные отклонения по главным осям 1 и 2, математические ожидания для классов -1 и 1, главная ось повернута на угол $\pi/4$ (см. рис. 2);
2) нормальные распределения с «шумом». К предыдущей модели добавлена «шумовая» компонента, для которой классы имеют одинаковые нормальные распределения с нулевыми средними и стандартным отклонением 5 по любому направлению. Вес (вероятность) компоненты равен 0,1;
3) модель с логистической функцией условной вероятности. Безусловное распределение $\mathsf{P}(dx)$ выбрано равномерным внутри эллипса с осями 2 и 1, главная ось повернута на угол $\pi/4$. Условная вероятность имеет вид $g(x) = \sigma(x)$;
4) к предыдущей модели добавлена шумовая компонента, как в модели 2;

Усредненные вероятности ошибочной классификации на различных моделях

Вероятностная модель	Методы классификации				
	Байесовское решающее правило	ЛДФ	Логистическая регрессия	SVM, $\varkappa = 0$	SVM, $\varkappa = 0{,}2$
1. Нормальное распределение	0,216	0,235	0,237	0,241	0,259
2. Нормальное распределение, шум	0,244	0,313	0,309	0,305	0,326
3. Логистическая ($g(x) = \sigma(x)$)	0,345	0,380	0,382	0,389	0,428
4. Логистическая, шум	0,359	0,433	0,430	0,431	0,464
5. Логистическая, смесь распределений	0,320	0,365	0,352	0,364	0,414
6. Логистическая ($g(x) = \sigma(2{,}5x)$)	0,202	0,220	0,223	0,226	0,256
7. Правдоподобия для критерия SVM	0,207	0,228	0,232	0,233	0,258

5) безусловное распределение $\mathsf{P}(dx)$ есть смесь равномерного распределения внутри эллипса из модели 3 и нормального распределения со стандартным отклонением 5 (по любому направлению). Вес второй компоненты 0,1. Условная вероятность есть $g(x) = \sigma(x)$;
6) модель, как в варианте 3, за исключением того, что $g(x) = \sigma(2{,}5x)$;
7) модель задается формулой (2) при $w\sqrt{2} = (1, 1)$ и изображена на рис. 3, *б*.

В таблице приведены полученные методом статистического моделирования оценки математических ожиданий вероятностей ошибочной классификации $\mathcal{F}(c, Q)$ для следующих методов: Байесовское (оптимальное) решающее правило; линейный дискриминант Фишера (ЛДФ); логистическая регрессия; метод SVM с параметром $\varkappa = 0$; метод SVM с $\varkappa = 0{,}2$.

Заметим, что на самом деле метод SVM запускался не при нулевом $\varkappa$, а при при $\varkappa = 0{,}0001$. Однако говорить о нулевом значении параметра $\varkappa$ все же допустимо, поскольку существует предел решения при $\varkappa \to 0$, и достаточно малые значения $\varkappa$ можно практически отождествлять с нулем.

Погрешность (в смысле стандартного отклонения) приведенных в таблице значений около 0,002.

Результаты приведены для объема выборки $N = 30$. Данное значение выбрано эмпирически как объем выборки, при котором различие методов проявляется в наибольшей степени. При других значениях N результаты качественно согласуются с приведенными.

Анализируя таблицу, можно заметить, что дискриминант Фишера оказался лучшим методом на моделях 1, 3, 6 и 7. Превосходство этого метода на модели 1 соответствует общепринятым ожиданиям (хотя убедительные обоснования для таких ожиданий отсутствуют), согласно которым ЛДФ принято ассоциировать с нормальными распределениями. Однако модель 6 существенно отличается от нормальной, но ЛДФ и на ней существенно лучше остальных методов.

Вместе с тем, на всех моделях, состоящих из смеси распределений, метод ЛДФ существенно проигрывает. Причина этого, очевидно, в том, что критерий дискриминанта Фишера содержит квадраты отклонений выборочных точек и поэтому неустойчив к редким большим отклонениям («выбросам»).

В целом, можно сделать вывод, что эффективность ЛДФ связана не с нормальностью распределений, а с наличием «выбросов». Действительно, в рамках исследования ЛДФ оказался лучшим на всех моделях без «выбросов».

Логистическая регрессия оказалась лучшей на большинстве оставшихся моделей, особенно на модели 5. Это вполне объяснимо, поскольку в этой модели функция условной вероятности имеет вид логистической кривой и при этом имеют место «выбросы».

В моделях 2 и 4 вид «шумовой» компоненты нарушает логистическую форму кривой условной вероятности, а наличие «выбросов» не позволяет получать хорошие решения методом ЛДФ. В результате, на модели 2 лучшим оказывается метод SVM.

Что касается параметра регуляризации, то во всех случаях решение при $\varkappa = 0$ оказывалось лучше, чем при $\varkappa = 0{,}2$. Вероятно, это связано как раз с тем, что ненулевые значения $\varkappa$ увеличивают вероятность того, что SVM не разобъет выборку, а отнесет ее к одному классу. Но все вероятностные модели таковы, что вероятность ошибочной классификации для такого решения равна 0,5.

Как и следовало ожидать, результаты логистической регрессии и SVM отличаются в большинстве случаев незначительно. Это объясняется тем, что методы различаются лишь эмпирическими функциями потерь, которые при этом достаточно близки.

7 Заключение

В работе поднята проблема построения вероятностных моделей, позволяющих выявлять свойства методов построения решающих функций и проводить исследование этих методов. В частности, ставилась задача построения моделей, на которых заданный метод наиболее эффективен среди сравиваемых методов.

Для метода логистической регрессии были построены модели, на которых этот метод эквивалентен ММП. Для метода SVM построена модель, на которой этот метод приближенно эквивалентен ММП. Для дискриминанта Фишера подобной модели построить не удалось.

Проблема построения набора «эталонных» вероятностных моделей для исследования и сравнения методов построения решающих функций остается практически полностью открытой. Вместе с тем, проведенное исследование демонстрирует принципиальную возможность продвижения в ее решении.

Также в работе выявлены некоторые неочевидные свойства метода SVM и особенности его поведения, учет которых позволяет более эффективно применять данный метод.

В работе были, в частности, установлены следующие любопытные факты:

- при применении метода SVM к модели нормальных распределений с равными матрицами величина зазора (ширина разделяющей полосы) уменьшается при удалении распределений друг от друга;
- существуют вероятностные модели, для которых Байесовская разделяющая функция линейна, но решения, полученные (на распределениях) методами SVM, логистической регрессии и дискриминанта Фишера не являются оптимальными;
- модель нормальных распределений с равными ковариационными матрицами не является моделью максимального правдоподобия для дискриминанта Фишера;
- дискриминант Фишера превосходит методы SVM и логистической регрессии на многих моделях с распределениями, далекими от нормального.

Полученные результаты позволяют лучше понять особенности исследуемых методов, что дает возможности для их дальнейшего усовершенствования.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Сравнение алгоритмов распознавания с помощью программной системы «Полигон» // Анализ данных и знаний в экспертных системах. — Новосибирск: Вычислительные системы, 1990. Вып. 134. С. 56–66.

[2] *Неделько В.М.* Регрессионные модели в задаче классификации // Сиб. ж. индустриальной математики, 2014. Т. XVII. № 1. С. 86–98.

[3] *Mease D., Wyner A.* Evidence contrary to the statistical view of boosting // J. Mach. Learn. Res., 2008. Vol. 9. P. 131–156.

[4] *Krasotkina O. V., Mottl V. V., Turkov P. A.* Bayesian aproach to the pattern recognition problem in nonstationary environment // Pattern recognition and machine intelligence / Eds. S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, S. K. Pal. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2011. Vol. 6744. P. 24–29.

[5] *Nedel'ko V. M.* Misclassification probability estimations for linear decision functions // Structural, syntactic, and statistical pattern recognition / Eds. A. Fred, T. M. Caelli, R. P. W. Duin, *et al.* — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2004. Vol. 3138. P. 780–787.

[6] *Неделько В. М.* К вопросу об эффективности бустинга в задаче классификации // Вестник Новосибирского гос. ун-та. Серия: математика, механика, информатика, 2015. Т. 15. Вып. 2. С. 72–89.

[7] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Институт математики СО РАН, 1999. 211 с.

[8] *Nedel'ko V.* Decision trees capacity and probability of misclassification // Autonomous intelligent systems: Agents and data mining / Eds. V. Gorodetsky, J. Liu, V. A. Skormin. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2005. Vol. 3505. P. 193–199.

[9] *Кельманов А. В., Пяткин А. В.* NP-трудность некоторых квадратичных евклидовых задач 2-кластеризации // Докл. РАН, 2015. Т. 464. № 5. С. 535–538.

[10] *Смердов С. О., Витяев Е. Е.* Синтез логики, вероятности и обучения: формализация предсказания // Сиб. электронные математические известия, 2009. Т. 6. С. 340–365.

[11] *Torshin I. Yu., Rudakov K. V.* On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification // Pattern Recogn. Image Anal., 2015. Vol. 25. No. 4. P. 577–587.

[12] *Лисицын Д. В.* Комбинированные регрессионные модели для описания данных, представленных в разных шкалах // Сб. научн. тр. Новосибирского гос. техн. ун-та, 2013. № 3(73). С. 41–48.

[13] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recogn. Image Anal., 2010. Vol. 20. No. 3. P. 269–285.

[14] *Motrenko A., Strijov V., Weber G.-W.* Sample size determination for logistic regression // J. Comput. Appl. Math., 2014. Vol. 255. P. 743–752.

[15] *Friedman J., Hastie T., Tibshirani R.* Additive logistic regression: A statistical view of boosting // Ann. Stat., 2000. Vol. 28. No. 2. P. 337–407.

[16] *Красоткина О. В., Турков П. А., Моттль В. В.* Байесовская логистическая регрессия в задаче обучения распознаванию образов при смещении решающего правила // Изв. Тульского гос. ун-та. Технические науки, 2013. № 2. С. 177–187.

[17] *Zhu J., Hastie T.* Support vector machines, kernel logistic regression and boosting // Multiple classifier systems / Eds. F. Roli, J. Kittler. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer, 2002. Vol. 2364. P. 16–26.

[18] *Середин О. С., Моттль В. В.* Метод опорных объектов для обучения распознаванию образов в произвольных метрических пространствах // Изв. Тульского гос. ун-та. Естественные науки, 2015. № 4. С. 49–66.

[19] *Lugosi G., Vayatis N.* On the Bayes-risk consistency of regularized boosting methods // Ann. Stat., 2004. Vol. 32. No. 1. P. 30–55.

[20] *Liu Y.* Fisher consistency of multicategory support vector machines // 11th Conference (International) on Artificial Intelligence and Statistics Proceedings. — San Juan, Puerto Rico, 2007. Vol. 2. P. 291–298.

[21] *Muandet K., Fukumizu K., Dinuzzo F., Schölkopf B.* Learning from distributions via support measure machines // Advances in neural information processing systems 25 / Eds. F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger. — MIT Press, 2012. P. 10–18.

[22] *Steinwart I.* Consistency of spport vector machines and other regularized kernel classifiers // IEEE Trans. Inform. Theory, 2005. Vol. 51. No. 1. P. 128–142.

[23] *Sollich P.* Bayesian methods for support vector machines: Evidence and predictive class probabilities // Mach. Learn., 2002. Vol. 46. P. 21–52.

[24] *Vojtěch F., Zien A., Schölkopf B.* Support vector machines as probabilistic models // Conference (International) on Machine Learning Proceedings. — New York, NY, USA: ACM, 2011. 665–672.

[25] *Боровков А. А.* О задаче распознавания образов // Теория вероятностей и её применение, 1971. Т. 16. № 1. С. 132–136.

[26] *Seredin O., Mottl V., Tatarchuk A., Razin N., Windridge D.* Convex support and relevance vector machines for selective multimodal pattern recognition // 21st Conference (International) on Pattern Recognition Proceedings. — Tsukuba, Japan, 2012. P. 1647–1650.

[27] *Татарчук А. И.* Байесовские методы опорных векторов для обучения распознаванию образов с управляемой селективностью отбора признаков. Дисс. ... канд. физ.-мат. наук, 2014. 125 с.

[28] *Salazar D. A., Vélez J. I., Salazar J. C.* Comparison between SVM and logistic regression: Which one is better to discriminate? // Rev. Colomb. Estad., 2012. Vol. 35. No. 2. P. 223–237.

Поступила в редакцию 31.08.2016

Investigation of effectiveness of several linear classifiers by using synthetic distributions*

V. M. Nedel'ko

nedelko@math.nsc.ru

S. L. Sobolev Institute of Mathematics SB RAS, 4 Acad. Koptyug Ave., Novosibirsk, Russia

The most common way to compare the effectiveness of data analysis methods is testing on tasks from UCI repository. However, this approach has several disadvantages, in particular, the incompleteness of the set of tasks and limited sample sizes. The present authors consider the possibility of building a repository of probabilistic distributions. The distributions are constructed purposefully in such a way as to reveal properties of the studied methods. Such distributions are called the probabilistic models. Some linear classification methods have been chosen for research: logistic regression, Fisher discriminant, and support vector machine. Several probabilistic models have been constructed to investigate the properties of these methods,

*The research was supported by the Russian Foundation for Basic Research (grants 14-01-00590 and 14-07-00249).

in particular, for each method, there was built a model on which this method outperformed the other methods. In addition, these models allow one to explain why a particular method was the best.

Keywords: *pattern recognition; machine learning; support vector machine; deciding function; logistic regression; misclassification probability*

DOI: 10.21469/22233792.2.3.04

References

[1] Lbov, G. S., and N. G. Starceva. 1990. Sravnenie algoritmov raspoznavaniya s pomoshch'yu programmnoy sistemy "Poligon" [Comparison of recognition algorithms with the software system "Poligon"]. *Analiz dannykh i znaniy v ekspertnykh sistemakh. Vychislitel'nye sistemy* [Analysis of data and knowledge in expert systems. Computer systems]. Novosibirsk. 34:56–66.

[2] Nedel'ko, V. M. 2014. Regressionnye modeli v zadache klassifikatsii [Regression models in the classification problem]. *Sib. zh. industrial'noy matematiki* [Siberian J. Industrial Mathematics] XVII(1):86–98.

[3] Mease, D., and A. Wyner. 2008. Evidence contrary to the statistical view of boosting. *J. Mach. Learn. Res.* 9:131–156.

[4] Krasotkina, O. V., V. V. Mottl, and P. A. Turkov. 2011. Bayesian approach to the pattern recognition problem in nonstationary environment. *Pattern recognition and machine intelligence*. Eds. S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, and S. K. Pal. Lecture notes in computer science ser. Berlin–Heidelberg: Springer-Verlag. 6744:24–29.

[5] Nedel'ko, V. M. 2004. Misclassification probability estimations for linear decision functions. *Structural, syntactic, and statistical pattern recognition*. Eds. A. Fred, T. M. Caelli, R. P. W. Duin, *et al.* Lecture notes in computer science ser. Berlin–Heidelberg: Springer-Verlag. 3138:780–787.

[6] Nedel'ko, V. M. 2015. K voprosu ob effektivnosti bustinga v zadache klassifikatsii [On the boosting efficiency in the classification problem]. *Vestnik Novosibirskogo gos. un-ta. Seriya: Matematika, mekhanika, informatika* [Bull. Novosibirsk State University. Ser. mathematics, mechanics, computer science] 15(2):72–89.

[7] Lbov, G. S., and N. G. Starceva. 1999. *Logicheskie reshayushchie funktsii i voprosy statisticheskoy ustoychivosti resheniy* [Logical decision functions and problem of statistical robustness of the solutions]. Novosibirsk: Institute of Mathematics SB RAS. 211 p.

[8] Nedel'ko, V. 2005. Decision trees capacity and probability of misclassification. *Autonomous intelligent systems: Agents and data mining*. Eds. V. Gorodetsky, J. Liu, and V. A. Skormin. Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag. 3505:193–199.

[9] Kel'manov, A. V., and A. V. Pyatkin. 2015. NP-hardness of some Quadratic Euclidean 2-clastering problems. *Dokl. Math.* 92(2):634–637.

[10] Smerdov, S. O., and E. E. Vityaev. 2009. Sintez logiki, veroyatnosti i obucheniya: Formalizatsiya predskazaniya [Probability, logic and learning synthesis: Formalizing prediction concept]. *Sibirskie Elektronnye Matematicheskie Izvestiya* [Siberian Electronic Math. Rep.] 6:340–365.

[11] Torshin, I. Yu., and K. V. Rudakov. 2015. On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification. *Pattern Recogn. Image Anal.* 25(4):577–587.

[12] Lisitsin, D. V. 2013. Kombinirovannye regressionnye modeli dlya opisaniya dannykh, predstavlennykh v raznykh shkalakh [Combined regression models for the data represented in different scales]. *Sb. nauchn. tr. Novosibirskogo gos. tekhn. un-ta* [Contributions of the Novosibirsk State Technical University] 3(73):41–48.

[13] Vorontsov, K. V. 2010. Exact combinatorial bounds on the probability of overfitting for empirical risk minimization. *Pattern Recogn. Image Anal.* 20(3):269–285.

[14] Motrenko, A., V. Strijov, and G.-W. Weber. 2014. Sample size determination for logistic regression. *J. Comput. Appl. Math.* 255:743–752.

[15] Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* 28(2):337–407.

[16] Krasotkina, O. V., P. A. Turkov, and V. V. Mottl'. 2013. Bayesovskaya logisticheskaya regressiya v zadache obucheniya raspoznavaniyu obrazov pri smeshchenii reshayushchego pravila [Bayesian logistic regression in the problem of pattern recognition learning on shifting decision rule] // *Izv. Tul'skogo gos. un-ta. Tehnicheskie nauki* [Proceedings of the Tula State University. Engineering] 2:177–187.

[17] Zhu, J., and T. Hastie. 2002. Support vector machines, kernel logistic regression and boosting. *Multiple classifier systems*. Eds. F. Roli and J. Kittler. Lecture notes in computer science ser. Berlin–Heidelberg: Springer. 2364:16–26.

[18] Seredin, O. S., and V. V. Mottl'. 2015. Metod opornykh ob"ektov dlya obucheniya raspoznavaniyu obrazov v proizvol'nykh metricheskikh prostranstvakh [The method of support objects for pattern recognition in arbitrary metric spaces]. *Izv. Tul'skogo gos. un-ta. Estestvennye nauki* [Proceedings of the Tula State University. Natural Sciences] 4:49–66.

[19] Lugosi, G., and N. Vayatis. 2004. On the Bayes-risk consistency of regularized boosting methods. *Ann. Stat.* 32(1):30–55.

[20] Liu, Y. 2007. Fisher consistency of multicategory support vector machines. *11th Conference (International) on Artificial Intelligence and Statistics Proceedings*. San Juan, Puerto Rico. 2:291–298.

[21] Muandet, K., K. Fukumizu, F. Dinuzzo, and B. Schölkopf. 2012. Learning from distributions via support measure machines. *Advances in neural information processing systems 25*. Eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. MIT Press. 10–18.

[22] Steinwart, I. 2005. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory* 51(1):128–142.

[23] Sollich, P. 2002. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Mach. Learn.* 46:21–52.

[24] Vojtěch, F., A. Zien, and B. Schölkopf. 2011. Support vector machines as probabilistic models. *Conference (International) on Machine Learning Proceedings*. New York, NY: ACM. 665–672.

[25] Borovkov, A. A. 1971. On the problem of pattern recognition. *Theor. Probab. Appl.* 16(1):141–144.

[26] Seredin, O., V. Mottl, A. Tatarchuk, N. Razin, and D. Windridge. 2012. Convex support and relevance vector machines for selective multimodal pattern recognition. *21st Conference (International) on Pattern Recognition Proceedings*. Tsukuba, Japan. 1647–1650.

[27] Tatarchuk, A. I. 2014. Bayesovskie metody opornykh vektorov dlya obucheniya raspoznavaniyu obrazov s upravlyaemoy selektivnost'yu otbora priznakov [Bayesian methods of support vector machines for pattern recognition training with controlled selectivity feature selection]. PhD Thesis. 125 p.

[28] Salazar, D. A., J. I. Vélez, and J. C. Salazar. 2012. Comparison between SVM and logistic regression: Which one is better to discriminate? // *Rev. Colomb. Estad.* 35(2):223–237.

Received August 31, 2016

Новый метод интеллектуального анализа и распознавания трехмерных изображений: описание и примеры*

Н. Г. Федотов[1]***, А. А. Сёмов***[2]***, А. В. Моисеев***[3]
fedotov@pnzgu.ru; mathematik_aleksey@mail.ru; moigus@mail.ru
[1]Пензенский государственный университет, Россия, г. Пенза, ул. Красная, д. 40
[2]ООО «Комэрф», Россия, г. Пенза, ул. Гагарина, д. 16
[3]Пензенский государственный технологический университет
Россия, г. Пенза, проезд Байдукова, ул. Гагарина, д. 1, а/11

Предлагается новый подход к распознаванию трехмерных (3D) объектов. Приведено подробное математическое описание метода, разработанного на основе указанного выше подхода. Описывается техника сканирования гипертрейс-преобразования и обосновывается выбор сканирующего элемента. Анализируются принципы интеллектуального анализа и распознавания 3D изображений, построенные на его основе. Предлагаемый метод основан на элементах стохастической геометрии и функционального анализа. Гипертрейс-преобразование обладает рядом преимуществ и возможностями интеллектуального анализа данных. Например, одной из интеллектуальных способностей предлагаемого метода является конструирование гипертриплетных признаков разной структуры («длинные» и «короткие» признаки). Разные типы признаков находят свое применение в принципах интеллектуального анализа и распознавания 3D изображений (верифицируемость и фальсифицируемость изображений). Ввиду только теоретического и концептуального характера статьи практические результаты не приводятся. Дается описание теоретических примеров построения «длинных» и «коротких» признаков изображений. Обосновывается их различие и особенности практического применения. Гипертрейс-преобразование имеет уникальную способность, аналогичную возможности человеческой зрительной системы, когда при достаточно беглом взгляде человек может быстро отличить друг от друга два пространственных объекта Данное обстоятельство повышает скорость работы сканирующей системы и надежность всей системы распознавания изображений в целом, улучшая интеллектуальные способности гипертрейс-преобразования.

Ключевые слова: *гипертрейс-преобразование; интеллектуальный анализ и распознавание 3D изображений; инвариантное описание; аналитическая структура гипертриплетного признака*

DOI: 10.21469/22233792.2.3.05

1 Введение

Устойчивой тенденцией научно-технического прогресса является увеличение числа людей, занятых обработкой информации, которое становится все больше и больше с каждым годом. Одной из важнейших задач, возникающих при создании информационных систем, является автоматизация процесса распознавания образов. Для ее решения ведутся широкие исследования, которые призваны помочь познать одно из основных свойств человеческого мозга — способность распознавать [2–1]. Для этого разрабатываются и создаются решающие предпосылки для построения интеллектуальных систем.

*Работа выполнена при финансовой поддержке РФФИ, проект № 15-07-04484.

Все подходы к анализу и распознаванию 3D изображений можно разделить на две большие условные группы: методы, которые требуют предварительной нормализации положения 3D объекта, и методы, которые дают инвариантное описание 3D объекта вне зависимости от его пространственного положения и ориентации. К современным исследованиям первой группы методов можно отнести работу [4]. Так, в данном исследовании предполагается, что для каждого 3D изображения непосредственно перед его распознаванием формируется карта глубины под разными углами обзора. Рассчитав карту глубины для текущего 3D изображения, находится наиболее близкий к этой карте глубины аналог из построенной базы карт. Зная углы обзора, по которым строилась база карт, определяют параметры вращения пространственного объекта, и строится соответствующая кубическая воксельная 3D сетка, для каждой ячейки которой рассчитываются различные признаки.

Недостаток данного метода состоит в том, что требуется не менее сотни углов обзора для получения множества синтетических карт глубины, чтобы эффективность распознавания была приемлемой. Кроме того, учитывается глубина пространственного объекта только под заданными углами обзора, а не его 3D форма в целом, вследствие чего, например, нельзя извлечь его геометрические характеристики.

Методы, которые используют плавающее окно 3D детектора обнаружения объекта и его границ, можно найти, например, в работах [5, 6].

Ко второй группе методов можно отнести работу [7]. Предложенный в ней метод относится к классу спектральных методов на стыке областей спектральной геометрии и дифференциальных уравнений. Этот метод является весьма перспективным, поскольку обеспечивает естественную библиотеку инструментов для анализа непосредственно поверхности 3D объекта в целом, а не его проекций. Данный метод позволяет обнаруживать повторяющиеся регионы на поверхности тела. Собственные функции оператора Лапласа–Бельтрами дают набор вещественных функций, которые предоставляют информацию о структуре и морфологии формы.

К очевидным недостаткам метода можно отнести тот факт, что признаки не имеют явной геометрической интерпретации и указывают лишь обобщенные свойства поверхности 3D объекта. Кроме того, данный метод, как пишут сами авторы, имеет две существенные проблемы: обнаружение адекватных регулярных областей поверхности 3D форм, обладающих одинаковыми свойствами кривизны, и согласование данных участков между собой (выделение четких границ).

Существуют также и другие методы, аналогичные выше описанным, которые концентрируются на анализе поверхности 3D объекта, но при этом используются другие операторы при конструировании дескриптора признаков. Так, дескрипторы с использованием 3D дискретного преобразования косинуса, применяемые для поиска 3D объектов в базах данных (аналогично двумерному аналогу, используемому в алгоритме сжатия JPEG), описаны в работе [8], дескрипторы с использованием анизотропной диффузии тензорных полей для анализа геометрии сгибов и деформаций анатомических органов и частей тела человека — в работе [9], с использованием семейства параметрических спектральных дескрипторов Лапласа для анализа и распознавания 3D человеческих фигур — в работе [10]. В целом, они обладают приблизительно теми же достоинствами и недостатками, что и описанная выше работа [7].

К категории современных методов, дающих инвариантное описание 3D изображения вне зависимости от его пространственного положения и ориентации, можно отнести также работу [11]. Данный метод заключается в извлечении инвариантного к группе движений

дескриптора признаков с использованием многомерной регрессионной линейной модели, примененной к описанию пространственного объекта, заданного в виде облака точек.

В последнее время получили развитие методы, основанные на представлении 3D изображений и точечных полей в виде кватернионных сигналов, предполагающие переход к формированию описания объекта в виде контура многогранника [12].

Основным недостатком всех рассмотренных методов является отсутствие возможности конструирования признаков, которые способны описывать форму и структуру объекта и вычислять его метрические характеристики. Вследствие слабых разработанных структур признаковых систем у различных методов заметно снижаются интеллектуальные возможности распознавания 3D изображений.

Подражание деятельности человеческого мозга — не единственный выход и подход к построению подобных систем, обладающей высокой интеллектуальной способностью [13]. У техники есть свои собственные пути реализации этой задачи, отличающиеся в техническом плане от естественных способностей человека, но учитывающие преимущества и особенности естественных физических зрительных систем. Раскрытию некоторых из этих путей, связанных с применением стохастической геометрии и функционального анализа, и посвящена настоящая статья.

Ниже приведено описание математической структуры метода с подробными комментариями по мере необходимости.

2 Техника сканирования трехмерного изображения и выбор сканирующего элемента

Пусть F — исходная модель 3D изображения. Определим плоскость $B(\eta, r) = \{ x | x^T \cdot \eta = r \}$ как касательную к сфере с центром в начале координат и с радиусом r, проходящую через заданную точку X и на расстоянии r от начала координат с заданными углами ω и φ, где $\eta = [\cos\omega \cdot \sin\varphi \sin\omega \cdot \sin\varphi \cos\varphi]$ — единичный вектор в $\mathbb{R}^3$, ω — угол между осью $0x$ и проекцией отрезка OX на плоскость $0xy$, φ — угол между осью $0z$ и отрезком OX (рис. 1).

Сканирование 3D изображения будет осуществляться плоскостями. Данный выбор оправдан тем, что пересечение плоскости с любым другим пространственным геометрическим примитивом хорошо известно в математике и имеет строгое аналитическое представление, которое пригодится при разработке аналитической структуры признака.

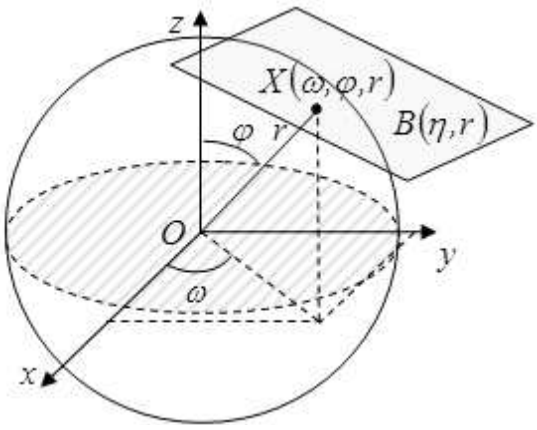
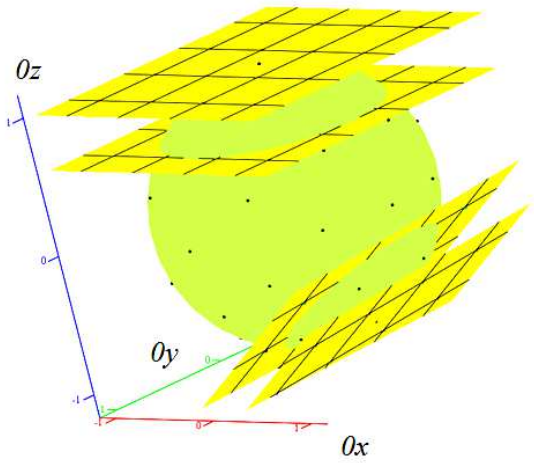
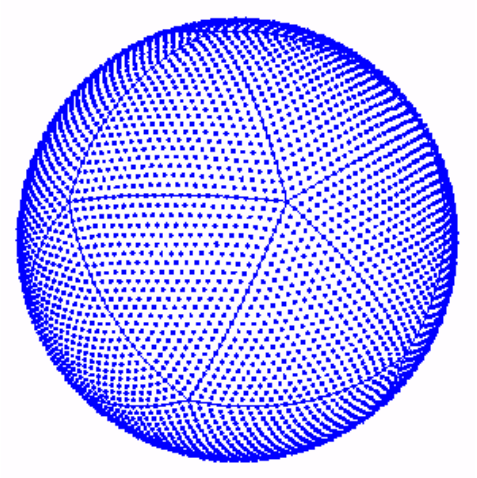


Рис. 1 Определение сферических координат плоскости



(*а*) Опорная сетка на сфере и соответствующие ей сетки сканирующих параллельных плоскостей



(*б*) Пример равномерной сетки на сфере

Рис. 2 Формирование опорной сетки на сфере

Кроме того, наличие сетки параллельных плоскостей помогает решить проблему инвариантного описания объекта (дает необходимое условие для конструирования признаков, инвариантных к группе движений 3D изображения). Так, если пространственный объект сканируется сеткой параллельных плоскостей, то перемещение исходного 3D изображения на любое расстояние вдоль прямой, содержащей вектор нормали сканирующей плоскости, не изменяет форму получаемых сканирующими плоскостями сечений (дискретный шаг сканирования игнорируется). Так как сканирование будет осуществляться под разными углами наклона плоскостей в пространстве для обзора пространственного объекта со всех сторон, то получаемые сечения и извлекаемые на их основе признаки не изменят своего значения при переносе 3D изображения на любой вектор в пространстве. В результате вычисляемые признаки не будут зависеть от пространственного положения объекта и его пространственной ориентации.

Таким образом, чтобы схема сканирования 3D изображения не была привязана к пространственной ориентации объекта, необходимо и достаточно, чтобы сканирующие элементы, если все их одновременно зафиксировать в пространстве, давали одинаковые сечения 3D объекта при любом его угле вращения. Другими словами, необходимо добиться, чтобы все сканирующие сетки параллельных плоскостей под разными углами ω и φ обзора распознаваемого трехмерного изображения совпадали бы друг с другом при любом его пространственном повороте (дискретный шаг игнорируется).

Стандартный перебор всех углов ω и φ, которыми идентифицируется каждая сканирующая сетка параллельных плоскостей, в топологическом смысле для непрерывного случая дает модель концентрических сфер с центром в начале координат. Каждой сканирующей сетке параллельных плоскостей на единичной сфере сопоставим точку, которая будет являться точкой касания со сферой плоскости, параллельной плоскостям данной сетки (отдельно для каждой пары (ω, φ) углов обзора). Множество точек на сфере образуют сетку, которую будем называть опорной (рис. 2).

Стоит отметить, что пара углов (ω, φ) однозначно определяет узел опорной сетки, соответствующий единственной касательной плоскости к сфере в этой точке, а значит, и единственной сетке сканирующих параллельных плоскостей.

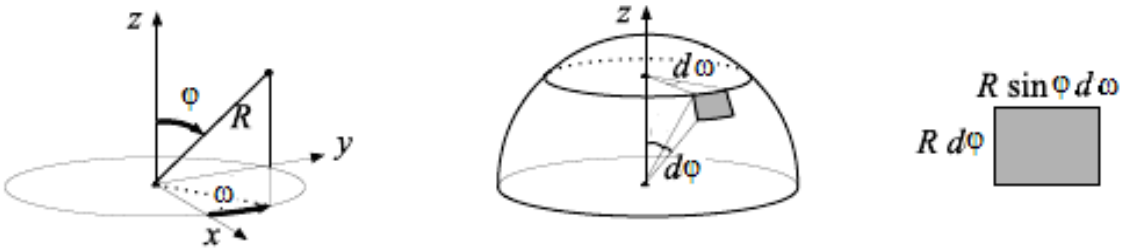


Рис. 3 Участок на поверхности сферы и значение его сторон

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 исходной и повернутой сеток. Так как каждая точка опорной сетки на сфере однозначно определяет угол наклона сетки параллельных плоскостей, то изменение угла наклона сетки плоскостей повлияет на форму получаемых сечений. Вследствие этого увеличится ошибка расчета признака и снизится точность распознавания 3D объекта.

С другой стороны, если при повороте сферы вокруг своего центра опорная сетка перейдет сама в себя, то соответствующие сетки секущих параллельных плоскостей полностью совпадут друг с другом и получаемые сечения будут одинаковыми (не изменят своей формы). Поэтому вычисляемое значение признака не изменится.

Таким образом, необходимо построить опорную сетку, обладающую равномерным распределением точек на сфере для достижения меньшей ошибки совмещения узлов опорной сетки при ее повороте из-за дискретного шага сканирования. Равномерное распределение точек опорной сетки на сфере (см. рис. 2) обеспечит отсутствие более плотных скоплений узлов опорной сетки на поверхности сферы, определяющих преимущественно сечения под теми или иными углами обзора объекта. В связи с этим, все результаты сканирования будут принимать равноправное участие при вычислении значения признака 3D изображения без повышения влияния каких-либо определенных значений сечений, так как частота появления любого среза сечений будет приблизительно одинакова (равномерный обзор 3D тела со всех сторон). Другими словами, значение вычисляемого признака не будет зависеть от ориентации 3D изображения в пространстве.

Указанное свойство равномерного распределения точек опорной сетки на сфере является необходимым условием инвариантности конструируемых признаков к повороту. Математическая формулировка данной проблемы имеет следующий вид.

Рассмотрим сферу, заданную в параметрическом виде: $x(\omega,\varphi) = R\cos\omega\cdot\sin\varphi$, $y(\omega,\varphi) = R\sin\omega\cdot\sin\varphi$, $z(\omega,\varphi) = R\cos\varphi$. Необходимо определить аналитически функцию $f(\omega,\varphi)$ плотности совместного распределения параметров ω и φ, соответствующую равномерному распределению точек на поверхности сферы.

Рассмотрим небольшой участок dS поверхности сферы, ограниченный приращениями $d\varphi$ и $d\omega$ (рис. 3).

В случае, когда точки имеют равномерное распределение на поверхности сферы, вероятность попадания произвольной точки A на элемент поверхности dS с одной стороны равна:

$$P(A\subset dS) = \frac{dS}{S}.$$

При постоянном значении φ изменение угла $d\omega$ описывает дугу $R\sin\varphi\, d\omega$. Поэтому площадь малого элемента поверхности сферы равна $dS = R^2 \sin\varphi\, d\omega d\varphi$. Таким образом, вероятность попадания произвольной точки A на элемент поверхности dS будет равна:

$$P\left(A \subset dS\right) = \frac{R^2 \sin\varphi\, d\omega d\varphi}{\int_0^{\pi}\int_0^{2\pi} R^2\, d\omega d\varphi} = \frac{R^2 \sin\varphi\, d\omega d\varphi}{4\pi R^2} = \frac{\sin\varphi\, d\omega d\varphi}{4\pi}.$$

С другой стороны, вероятность попадания точки A на данный элемент поверхности равна: $P\left(A \subset dS\right) = f\left(\omega, \varphi\right)\, d\omega d\varphi$. Следовательно, совместная плотность распределения вероятности ω и φ равна:

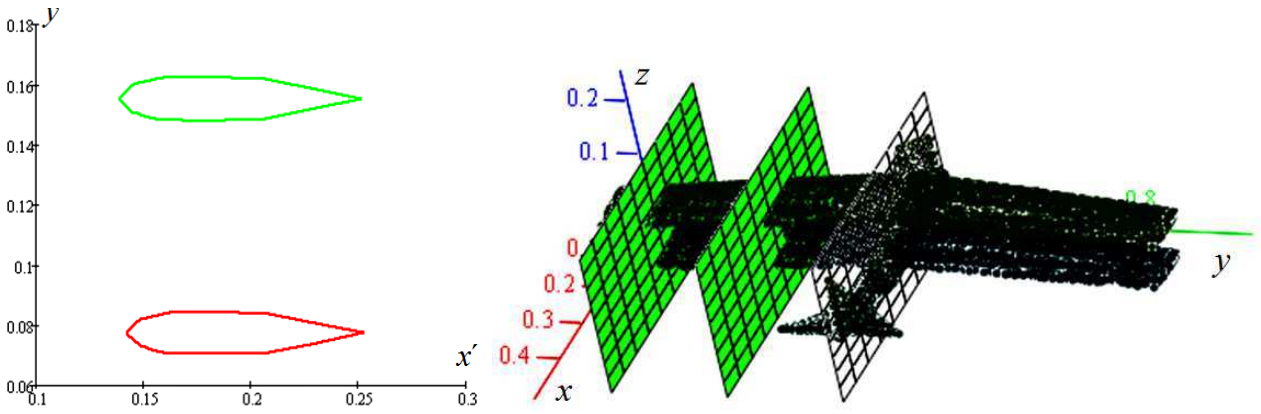
$$\frac{\sin\varphi\, d\omega d\varphi}{4\pi} = f\left(\omega, \varphi\right)\, d\omega d\varphi \Rightarrow f\left(\omega, \varphi\right) = \frac{\sin\varphi}{4\pi}.$$

При генерации значений параметров ω и φ с использованием функции $f\left(\omega, \varphi\right)$ будет получаться равномерное распределение точек на поверхности сферы. Более подробно о построении равномерных опорных сеток на сфере согласно технике сканирования предлагаемого метода можно найти в [14].

3 Математическая модель гипертрейс-преобразования

Сканирование 3D изображения производится сеткой параллельных плоскостей с расстоянием Δr между плоскостями и заданными углами ω и φ (рис. 4). Взаимное положение 3D изображения F и каждой сканирующей плоскости $B\left(\eta(\omega, \varphi), r\right)$ образует сечение (см. рис. 4), которое характеризуется числом G, определяемым по некоторому правилу HyperT : $G = \text{HyperT}\left(F \bigcap B\left(\eta(\omega, \varphi), r\right)\right)$. В качестве указанного правила можно использовать вычисление периметра или площади сечения, количество пересечений плоскости с исходным объектом, свойства окрестности полученного сечения и т. п. Другими словами, функционал HyperT характеризует свойство признака сечения.

Затем сканирование производится сеткой параллельных плоскостей для нового значения угла $\omega + \Delta\omega$ и $\varphi + \Delta\varphi$, получившего дискретные приращения $\Delta\omega$ и $\Delta\varphi$ соответственно, с тем же шагом Δr между сканирующими элементами сетки плоскостей. К сечениям новой сетки из плоскостей $B\left(\eta(\omega + \Delta\omega, \varphi + \Delta\varphi), r_i\right)$ применяется такое же ранее выбранное



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

Рис. 4 Особенности сканирования 3D объекта

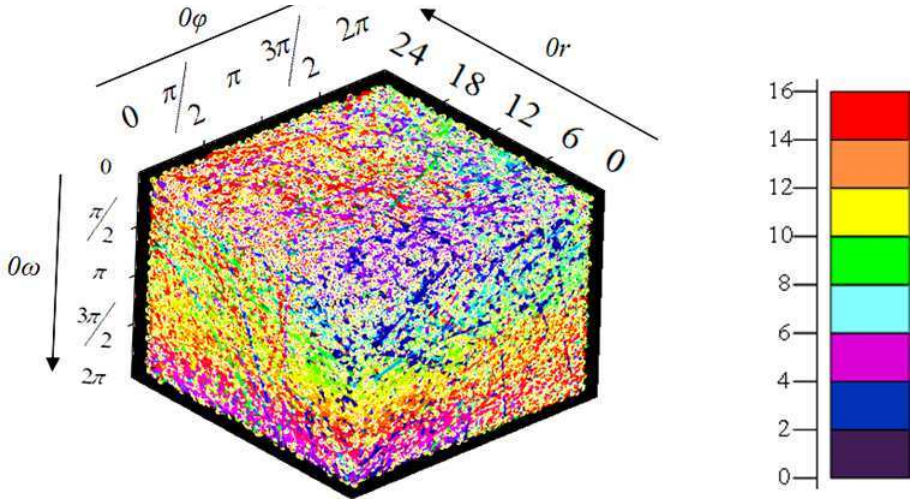


Рис. 5 Пример графического представления гипертрейс-матрицы 3TM

правило HyperT. Важно отметить, что углы изменяются не произвольным образом, а согласно построению опорной сетки на сфере.

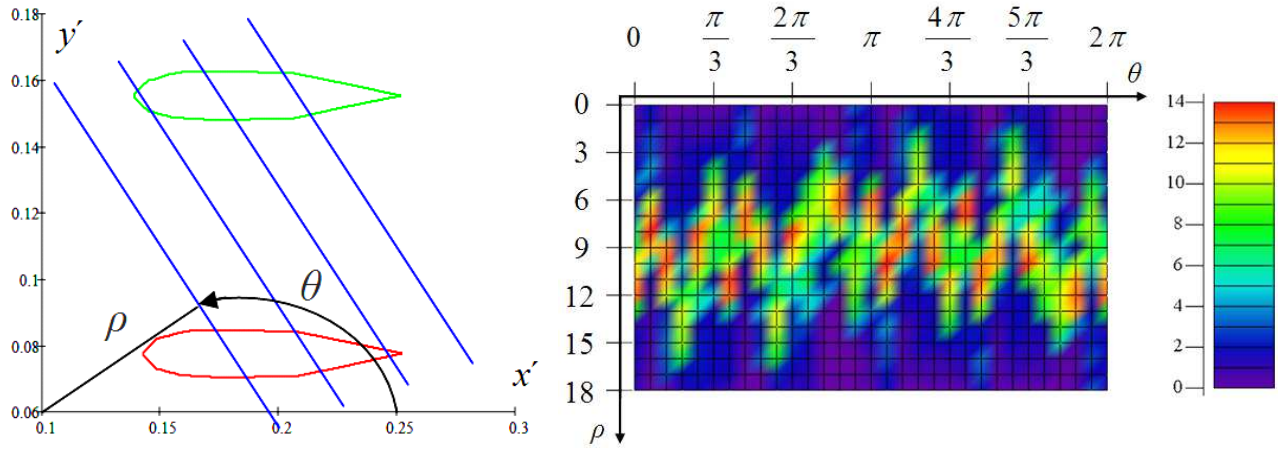
Результат вычислений функционала HyperT зависит от трех параметров плоскости $(r,\ \omega,\ \varphi)$. Поэтому если каждому двумерному (2D) изображению, полученному при сечении исходной 3D модели сканирующей плоскостью, сопоставить некоторый информативный признак $\Pi\left(F_{\text{sect}}\right)$ по правилу HyperT, то при численном анализе результат 3D трейс-преобразования удобно представить в виде 3D гипертрейс-матрицы 3TM, у которой ось 0φ направлена вертикально, ось 0ω — горизонтально, ось $0r$ — вглубь [15].

Например, каждая глубинная строка матрицы содержит элементы-признаки, которые вычисляются по 2D изображениям, полученным в результате сечений исходного 3D объекта сканирующими плоскостями при обходе всех значений переменной расстояния r с фиксированными значениями углов ω и φ. Если плоскость B не пересекает 3D изображение, т. е. $F \bigcap B\left(\eta(\omega,\varphi),r\right) = \varnothing$, то значение гипертрейс функционала полагают равным нулю: $\text{HyperT}\left(F \bigcap B\left(\eta(\omega,\varphi),r\right)\right) = 0$.

Графическое представление гипертрейс-матрицы 3TM называется гипертрейс трансформантой (рис. 5), где полученное в результате сканирования множество чисел G образует точки $(\omega_i,\ \varphi_j,\ r_k)$ в системе координат с осями 0ω, 0φ и $0r$. Стоит отметить, что в данном случае элемент матрицы показывает значение периметра соответствующего сечения.

Таким образом, тройке $(\omega_i,\ \varphi_j,\ r_k)$ соответствует элемент матрицы с номером $(i,\ j,\ k)$ и значением $\Pi\left(F_{\text{sect}}\right)$, который характеризует информативный признак 2D фигуры F_{sect}, полученной в сечении объекта F плоскостью $B\left(\eta(\omega_i,\varphi_j),r_k\right)$: $F_{\text{sect}} = F \bigcap B\left(\eta(\omega_i,\varphi_j),r_k\right)$. Так как результат вычислений функционала $\text{HyperT}\left(F \bigcap B\left(\eta(\omega_j,\varphi_i),r_k\right)\right)$ зависит от трех параметров сканирующей плоскости r, ω и φ, имеющих дискретный шаг сканирования, то реальная гипертрейс трансформанта имеет дискретную структуру.

После заполнения 3D гипертрейс-матрицы обрабатываются ее глубинные строки с помощью функционала HyperP, который можно задать, например, как $\text{HyperP} = \int G\left(\omega,\varphi,r\right) dr$. В результате исходная 3D гипертрейс-матрица 3TM становится двумерной матрицей 2TM. Далее применяется постолбцовая обработка матрицы 2TM посредством функционала HyperΩ, который можно задать, например, как $\text{HyperΩ} = \max\limits_{\varphi} G\left(\omega,\varphi\right)$.



(*а*) Процесс сканирования 2D сечения сеткой(*б*) Пример графического представления трейс-матрицы TM параллельных прямых

Рис. 6 Особенности сканирования 2D объекта

В результате получается горизонтальная строка 1TM — вектор значений, непрерывным аналогом которого является 2π-периодическая кривая. К полученному набору чисел применяют функционал HyperΘ, что приводит к появлению некоторого числа — признака изображения $\mathrm{Res}(F)$. Этот функционал можно задать, например, амплитудой второй гармоники ряда Фурье от дискретной функции значений элементов 1TM или любым другим функционалом (например, минимум от множества значений элементов строки 1TM).

Таким образом, гипертриплетный признак 3D изображения F обладает структурой в виде композиции четырех функционалов, каждый из которых кроме функционала HyperT при последовательном применении сокращает размерность матрицы 3TM на единицу [16]:

$$\mathrm{Res}\,(F) = \mathrm{HyperΘ} \circ \mathrm{HyperΩ} \circ \mathrm{HyperP} \circ \mathrm{HyperT}\,(F_{\mathrm{sect}})\,.$$

Каждое 2D изображение, получившееся в сечении исходной 3D модели сеткой параллельных плоскостей под разными углами обзора, необходимо просканировать, чтобы извлечь какие-нибудь значимые признаки (например, периметр контура фигуры сечения). Для нахождения признака 2D изображения сечения используется трейс-преобразование [17].

Сканирование получаемых в сечении фигур F_{sect} осуществляется решеткой параллельных прямых $l\,(\theta, \rho)$ с расстоянием $\Delta\rho$ между линиями, где ρ и θ — полярные координаты прямой в плоскости сечения (рис. 6). Взаимное положение 2D изображения F_{sect} и каждой сканирующей линии $l\,(\theta, \rho)$ характеризуется числом g, вычисляемым по некоторому правилу $\mathrm{T} : g = \mathrm{T}\,(F_{\mathrm{sect}} \bigcap l\,(\theta, \rho))$. В качестве указанного правила можно использовать вычисление длины части прямой, лежащей внутри изображения, свойства окрестности точки пересечения прямой с изображением и т. п.

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 $l\,(\theta + \Delta\theta, \rho)$ и сечения F_{sect} применяется такое же ранее выбранное правило T. Сканирование повторяется для каждого нового угла $\theta + \Delta\theta$ до завершения оборота в 2π радиан.

Результат вычислений функционала T зависит от двух параметров прямой θ и ρ. При численном анализе результат 2D трейс-преобразования удобно представить в виде 2D трейс-матрицы TM, у которой ось 0θ направлена горизонтально, а ось 0ρ — вертикально (см. рис. 6).

Например, каждый вертикальный столбец матрицы TM содержит значения, вычисляемые по всем прямым сканирующей сетки при одинаковом значении угла θ для одного и того же 2D изображения сечения. Если прямая l не пересекает изображение: $F_{\text{sect}} \bigcap l\,(\theta, \rho) = \varnothing$, то значение трейс-функционала полагают равным нулю $\mathrm{T}\,(F_{\text{sect}} \bigcap l\,(\theta, \rho)) = 0$.

Графическое представление трейс-матрицы TM называется трейс трансформантой, где полученное в результате сканирования множество чисел g образуют точки $(\theta_i,\, \rho_j)$ в системе координат с осями 0θ и 0ρ. Сам элемент матрицы показывает значение отрезка максимальной длины из множества отрезков, высекаемых одной сканирующей в 2D фигуре сечения.

Таким образом, паре $(\theta_i,\, \rho_j)$ соответствует элемент матрицы TM с номером (i, j) и значением $\mathrm{T}\,(F_{\text{sect}} \bigcap l\,(\theta_i, \rho_j))$.

После заполнения 2D трейс-матрицы с помощью диаметрального функционала P обрабатываются столбцы матрицы TM. Его можно задать, например, как $\mathrm{P} = \int g\,(\theta, \rho)\; d\rho / \max\limits_{\rho} g\,(\theta, \rho)$. В результате исходная 2D матрица TM становится одномерной матрицей — вектором чисел, непрерывным аналогом которого будет 2π-периодическая кривая. Затем к полученному набору чисел применяют функционал Θ, который можно задать как $\Theta = \min\limits_{\theta} g\,(\theta)$. В результате получается число $\Pi\,(F_{\text{sect}})$ — признак 2D изображения сечения F_{sect}.

Таким образом, триплетный признак 2D изображения F_{sect} обладает структурой в виде композиции трех функционалов, каждый из которых кроме функционала T при последовательном применении сокращает размерность матрицы TM на единицу [18]:

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

Объединяя полученные формулы для $\mathrm{Res}(F)$ и $\Pi\,(F_{\text{sect}})$, окончательно получаем следующую аналитическую структуру признака 3D изображения в виде композиции множества функционалов:

$$\mathrm{Res}\,(F) = \mathrm{Hyper}\Theta \circ \mathrm{Hyper}\Omega \circ \mathrm{HyperP} \circ \mathrm{HyperT}\left(\Theta \circ \mathrm{P} \circ \mathrm{T}\left(F_{\text{sect}} \bigcap l\,(\theta, \rho)\right)\right).$$

Благодаря композиционной структуре функционалов, входящих в аналитическую структуру признака $\Pi\,(F_{\text{sect}})$ и $\mathrm{Res}(F)$ соответственно, возможно получение огромного числа признаков. Причем некоторые признаки 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.

Подробное описание особенностей техники стохастического сканирования, ее преимущества и влияние на формирование признака можно найти в [19].

Стоит отметить, что равномерная сетка на сфере неизоморфна равномерной сетке на плоскости. Ввиду этого, при переходе от координат равномерной сетки на сфере к координатам элементов гипертрейс-матрицы 3TM возникают определенные трудности сохранения целостности ее структуры — нарушается порядок следования строк и столбцов

друг за другом в матрице. Так, при обработке функционалами 3D матрица сворачивается в число в строго заданных направлениях (сначала глубинные, далее вертикальные, а затем горизонтальные строки). Поэтому из-за произвольной неизвестной ориентации тела в пространстве возможен случайный поворот матрицы относительно оси $0r$ в пространстве $0\omega\varphi r$. Произвольное нарушение порядка следования строк матрицы друг за другом в данных фиксированных направлениям приведет к изменению значения вычисляемого признака, и, как следствие, инвариантность распознавания 3D изображения будет нарушена.

В связи с этим, правила нумерации узлов опорной сетки, по которым формируется 3D гипертрейс-матрица и определяется порядок следования ее элементов, должны определяться не относительно координатных осей, а относительно произвольно ориентированного пространственного объекта. Техника предлагаемого метода позволяет определить данную ориентацию достаточно просто в процессе сканирования объекта, не производя для этого дополнительного сканирования.

Для определения направления нумерации узлов опорной сетки достаточно идентифицировать некоторые опорные ключевые точки — узлы опорной сетки, которые однозначно определяются вне зависимости от пространственной ориентации 3D объекта. Данные опорные точки определяют начало отсчета (первую глубинную строку), от которого начинается заполняться гипертрейс-матрица. Например, построение гипертрейс-матрицы при нумерации узлов опорной сетки от первой ключевой точки по часовой стрелке в направлении второй ключевой точки для заполнения элементов матрицы.

Ключевые точки характеризуют уникальные свойства пространственного объекта, на основании которых могут быть построены гипертриплетные признаки, инвариантные к повороту 3D изображения. Так, в качестве ключевой точки, например, можно взять узел на опорной сетке, который соответствует сетке секущих плоскостей, содержащей максимальное по площади сечение исходного 3D объекта.

Стоит отметить, что при повороте объекта секущие плоскости (например, сечения, перпендикулярные главной оси объекта) будут соответствовать другим точкам опорной сетки, а не точкам сетки до его поворота. Поэтому необходимо задать правило, определяющее соотношение каждой точки опорной сетки смещенного объекта с соответствующей точкой опорной сетки исходного объекта. Другими словами, с использованием ключевых точек проблема перехода от координат элементов равномерной сетки к координатам 3D гипертрейс-матрицы транформируется в проблему создания правила инвариантной нумерации узлов опорной сетки на сферу.

Один из способов нумерации узлов опорной сетки заключается в следующем. Из всего множества возможных узлов опорной сетки выбирается ключевая точка, имеющая отличительное значение по какому-либо признаку. Данная ключевая точка считается за северный полюс, от которого по сфере начинаются строиться меридианы. Данные меридианы разбивают сферу на двуугольники с вершинами в северном и диаметрально противоположном ему относительно центра сферы южном полюсе. Нулевым меридианом будет считаться тот, который проходит через точку S с координатами $(1; 0; 0)$. В случае совпадения указанной точки S с одним из полюсов сферы — через точку $(0; 1; 0)$.

Нумерация в каждом двуугольнике идет от северного полюса к южному по часовой стрелке. Обход всех двуугольников совершается по часовой стрелке (если смотреть на северный полюс сверху), начиная и заканчивая двуугольниками, чьи стороны содержат нулевой меридиан.

В этом случае смысл строк и столбцов гипертрейс-матрицы 3ТМ останется тем же с той лишь разницей, что тройке $(\omega_i, \varphi_j, r_k)$ соответствует элемент матрицы со значением $\Pi\left(F_{\text{sect}}\right)$ и номером $(\delta_t, \gamma_s, r_k)$, соответствующая точка которого лежит в пределах от долготы δ_t до δ_{t+1}, в полосе широт от γ_s до γ_{s+1}. Соответственно, все точки, принадлежащие одному двуугольнику, соответствуют всем значениям одного вертикального столбца матрицы 3ТМ.

Горизонтальные строки определяют порядок считывания граней выпуклого многогранника (в частности, икосаэдра). Так как нулевой меридиан определяется на основе глобальной системы координат, от которой зависит ориентация 3D объекта, то вертикальные столбцы (ось 0δ) представляют собой дискретный аналог периодической кривой. При пространственном повороте объекта горизонтальные строки будут меняться на величину той части угла, которая влияет на поворот объекта вокруг оси полюсов. Таким образом, при повороте 3D трейс-образ изображения будет двигаться вдоль оси 0δ. Более подробно объяснение данной проблемы и пути ее решения можно найти в [20].

4 Построение гипертриплетных признаков разных категорий

Следует отметить, что гипертрейс-преобразование имеет уникальную способность, аналогичную возможности человеческой зрительной системы, когда при достаточно беглом взгляде человек может быстро отличить друг от друга два пространственных объекта. Данное свойство отчетливо видно при конструировании новых категорий 2D и 3D признаков, описание которых будет приведено ниже.

Пусть введена какая-либо мера расстояния $\rho\left(x, x'\right)$ между двумя объектами x и x'. «Сходство» 3D изображений между собой будет определяться функцией расстояния $\rho\left(\text{desk}\left(x\right), \text{desk}\left(x'\right)\right)$ между двумя векторами дескрипторов признаков образов $\text{desk}\left(x\right)$ в пространстве объектов X. Пример такой метрики и процедура определения класса 3D изображений, адаптированная под данный метод распознавания, представлены в [21].

Рассмотрим стандартные гипертриплетные признаки, описанные в разд. 3, которые имеют полную стандартную композиционную структуру функционалов (будем называть их «длинными»):

$$\text{Res}\left(F\right) = \text{Hyper}\Theta \circ \text{Hyper}\Omega \circ \text{HyperP} \circ \text{HyperT}\left(\Theta \circ \text{P} \circ \text{T}\left(F_{\text{sect}} \bigcap l\left(\theta, \rho\right)\right)\right).$$

Отметим, что функционалы HyperT и T отвечают за сканирование 3D и 2D изображений соответственно. Функционал HyperP, как и функционал P, отвечает за выполнение свойства инвариантности признаков к переносу изображения, необходимое условие которой достигается за счет использования сканирующих сеток параллельных плоскостей и прямых. Функционалы HyperΘ и HyperΩ, как и функционал Θ, оказывают влияние на выполнение свойства инвариантности к повороту, необходимое условие которого достигается за счет специфики техники сканирования — равномерное сканирование пространственного объекта со всех сторон.

Ниже приведен пример и описание «длинного» признака, который будет инвариантен к группе движений и масштабированию 3D изображения:

$$\text{Res}\left(F\right) = \text{Hyper}\Theta \circ \text{Hyper}\Omega \circ \text{HyperP} \circ \text{HyperT}\left(\Theta \circ \text{P} \circ \text{T}\left(F_{\text{sect}} \bigcap l\left(\theta, \rho\right)\right)\right),$$

где $\text{T}\left(F_{\text{sect}} \bigcap l(\theta, \rho)\right) = \min\limits_{t} f\left(\theta, \rho, t\right)$; $\text{P} = \sum_{\rho} g\left(\theta, \rho\right)$; $\Theta = \left(\max\limits_{\theta} g\left(\theta\right) + \min\limits_{\theta} g\left(\theta\right)\right) \Big/ 2$; $\text{HyperT}\left(F \bigcap B\left(\eta(\omega, \varphi), r\right)\right) = \Pi\left(F_{\text{sect}}\right) = G\left(\omega, \varphi, r\right)$; $\text{HyperP} = \max\limits_{r} G\left(\omega, \varphi, r\right)$; $\text{Hyper}\Omega = \text{Local}\max\limits_{\varphi} G\left(\omega, \varphi\right)$; $\text{Hyper}\Theta = \min\limits_{\omega} G\left(\omega\right)$; $f\left(\theta, \rho, t\right)$ — длина t-го отрезка, высекаемого ρ-й

прямой под θ-м углом в плоскости сечения F_{sect}; $\Pi\left(F_{sect}\right) = G\left(\omega, \varphi, r\right)$ — признак сечения, получаемого r-й плоскостью под парой углов (ω, φ) обзора 3D объекта.

Так, функционал T для каждой сканирующей прямой из сетки параллельных прямых находит минимальную длину отрезка, высекаемой одной прямой на 2D изображении сечения F_{sect}. Функционал P для каждой сетки параллельных прямых вычисляет сумму подсчитанных выше максимальных отрезков (отдельно для каждого угла наклона θ сетки параллельных прямых в плоскости сечения). Функционал Θ для всего множества сеток прямых под разными углами наклона θ вычисляет полусумму максимального и минимального значений среди подсчитанных выше сумм.

Далее функционал HyperT, используя полную структуру гипертриплетного и триплетного признаков, формирует гипертрейс-матрицу из вычисленных значений — признаков $\Pi\left(F_{\text{sect}}\right) = \Theta \circ \mathrm{P} \circ \mathrm{T}\left(F_{\text{sect}} \bigcap l(\theta, \rho)\right)$ сечений плоскостями исходного 3D изображения. Функционал HyperP для каждой сетки параллельных плоскостей вычисляет максимальное значение указанного признака сечений $\Pi\left(F_{\text{sect}}\right)$ (отдельно для каждой пары углов (ω, φ) обзора 3D изображения). Функционал HyperΩ вычисляет число локальных максимумов функции, образованной дискретным рядом элементов вертикальных строк (ось 0φ), содержащих подсчитанные выше значения максимального значения признака. Функционал HyperΘ среди подсчитанных выше значений числа локальных максимумов выбирает в получившейся строке минимальный элемент (ось 0ω).

На основе «длинных» признаков можно построить систему, позволяющую извлечь из всего объема подмножество информативных признаков, которое позволяет провести верификацию 3D изображения, что подтверждают работы [16, 21]. В силу этого признаки, включенные в это множество, можно условно назвать признаками верификации 3D изображения исходя из того, что если для двух изображений F и F' наблюдается близость векторов дескрипторов $\mathrm{desk}\left(F\right)$ и $\mathrm{desk}\left(F'\right)$, составленных из этих признаков, в смысле метрики $\rho\left(\mathrm{desk}\left(F\right), \mathrm{desk}\left(F'\right)\right)$, то изображения F и F' считаются одинаковыми или схожими, тогда как если соответствующая близость векторов дескрипторов признаков верификации («длинных» признаков) $\mathrm{desk}\left(F\right)$ и $\mathrm{desk}\left(F'\right)$ не наблюдается, то изображения F и F' считаются разными или непохожими (интеллектуальный принцип верификации 3D изображений).

Рассмотрим гипертриплетные признаки, которые имеют сокращенную композиционную структуру функционалов («короткие» признаки):

(1) $\mathrm{Res}\left(F\right) = \mathrm{HyperP} \circ \mathrm{HyperT}\left(\mathrm{P} \circ \mathrm{T}\left(F_{\text{sect}} \bigcap l\left(\theta, \rho\right)\right)\right)$;
(2) $\mathrm{Res}\left(F\right) = \mathrm{Hyper}\Theta \circ \mathrm{Hyper}\Omega \circ \mathrm{HyperT}\left(\Theta \circ \mathrm{T}\left(F_{\text{sect}} \bigcap l\left(\theta, \rho\right)\right)\right)$.

Как видно из приведенных выше формул, здесь опущены те или иные категории функционалов P, Θ, HyperP, HyperΩ и HyperΘ, которые отвечают за обработку результатов сканирования 3D изображений и получаемых плоских сечений и связаны таким образом с пространственным положением, ориентацией и уровнем масштабирования 3D изображения, поэтому для сохранения инвариантности распознавания 3D изображений к группе движений и масштабированию сокращенная аналитическая структура «коротких» признаков должна характеризовать свойства объекта, которые не изменяются при его масштабировании и движении. Для достижения данной цели нужно использовать функционалы, инвариантные к группе движений и масштабированию изображения, либо инвариантность распознавания достигать за счет комбинации отношения уже вычисленных признаков.

Выделим среди «коротких» признаков такие признаки, что если для двух изображений F и F' не наблюдается близость векторов дескрипторов $\mathrm{desk}\left(F\right)$ и $\mathrm{desk}\left(F'\right)$, состав-

ленных из этих признаков, в смысле метрики $\rho\left(\mathrm{desk}\left(F\right),\mathrm{desk}\left(F'\right)\right)$, то изображения F и F' считаются заведомо разными или непохожими, тогда как если соответствующая близость векторов дескрипторов выделенных «коротких» признаков $\mathrm{desk}\left(F\right)$ и $\mathrm{desk}\left(F'\right)$ наблюдается, то нельзя сделать никакой вывод о схожести или одинаковости изображений F и F' (интеллектуальный принцип фальсифицируемости 3D изображений). В силу вышесказанного, выделенные «короткие» признаки можно условно назвать признаками фальсификации 3D изображения.

К признакам фальсификации («коротким» признакам) можно отнести признаки, описывающие различные геометрические характеристики пространственного объекта (например, объем тела, аименьший радиус сферы, в которую можно поместить исходный пространственный объект и т. п.), а также признаки, описывающие различные свойства одиночной сканирующей прямой или плоскости (например, сечение с максимальной площадью сечения, наличие пустых полостей внутри 3D изображения и т. п.).

Ниже приведены некоторые конкретные примеры «коротких» признаков разных классов с описанием их аналитической структуры. Стоит отметить, что такие виды признаков могут иметь как явную геометрическую интерпретацию, так и неявные характеристики.

1. Признак пространственного объекта, не имеющий явной геометрической интерпретации:
$$\mathrm{Res}\left(F\right)=\mathrm{HyperP}\circ\mathrm{HyperT}\left(\mathrm{P}\circ\mathrm{T}\left(F_{\mathrm{sect}}\bigcap l\left(\theta^{*},\rho\right)\right)\right),$$
где $\mathrm{T}=\sum_{t}f\left(\theta=\theta^{*},\rho,t\right)$; $\mathrm{P}=\sum_{\rho}g\left(\theta^{*},\rho\right)$; $\mathrm{HyperT}=G\left(\omega=\omega^{*},\varphi=\varphi^{*},r\right)$; $\mathrm{HyperP}=\underset{r}{\mathrm{gmean}}\,G\left(\omega^{*},\varphi^{*},r\right)$; gmean — функция среднегармонического элементов вектора-строки G (вырожденные в единственный элемент вертикальные и горизонтальные строки матрицы G); ω^{*} и φ^{*} (или θ^{*}) означают, что сканирование осуществляется сеткой параллельных плоскостей (или прямых) только под одним углом наклона в пространстве (или плоскости сечения). Для большей точности и надежности распознавания можно производить сканирования под 2–3 различными случайными углами наклона, а полученные результаты усреднять. Например, $\mathrm{P}'=\sum_{i=1}^{3}\sum_{\rho}g\left(\theta_{i},\rho\right)/3$ и $\mathrm{HyperP}'=\sum_{j=1}^{2}\sum_{i=1}^{2}\sum_{r}G\left(\omega_{i},\varphi_{j},r\right)/4$.
2. Площадь поверхности пространственного объекта:
$$\mathrm{Res}\left(F\right)=\mathrm{HyperP}\circ\mathrm{HyperT}\left(\mathrm{P}\circ\mathrm{T}\left(F_{\mathrm{sect}}\bigcap l\left(\theta^{*},\rho\right)\right)\right),$$
где $\mathrm{T}=\sum_{t}f\left(\theta=\theta^{*},\rho,t\right)$; $\mathrm{P}=\left(\mathrm{Row2D}+1\right)\Delta\rho+\sum_{i=1}^{\mathrm{Row2D}-1}\left|g\left(\theta^{*},\rho_{i+1}\right)-g\left(\theta^{*},\rho_{i}\right)\right|$; $\mathrm{HyperT}=G\left(\omega=\omega^{*},\varphi=\varphi^{*},r\right)$; $\mathrm{HyperP}=\sum_{r}G\left(\omega^{*},\varphi^{*},r\right)\Delta r$; $f\left(\theta,\rho,t\right)$ — длина t-го отрезка, высекаемого ρ-й прямой под θ-м углом наклона в плоскости сечения F_{sect}; $\Delta\rho$ — расстояние между параллельными прямыми в плоскости сечения; Δr — расстояние между параллельными плоскостями в пространстве; Row2D — количество ненулевых элементов в столбце трейс-матрицы TM.
3. Максимальная длина отрезка, который может быть помещен внутри пространственного объекта:
$$\mathrm{Res}\left(F\right)=\mathrm{Hyper}\Theta\circ\mathrm{Hyper}\Omega\circ\mathrm{HyperT}\left(\Theta\circ\mathrm{T}\left(F_{\mathrm{sect}}\bigcap l\left(\theta,\rho^{*}\right)\right)\right),$$
где $\mathrm{T}=\underset{t}{\max}\left(f\left(\theta,\rho=\rho^{*},t\right)\right)$; $\Theta=\underset{\theta}{\max}\,g\left(\theta\right)$; $\mathrm{HyperT}=G\left(\omega,\varphi,r=r^{*}\right)$; $\mathrm{Hyper}\Omega=\underset{\varphi}{\max}\,G\left(\omega,\varphi\right)$; $\mathrm{Hyper}\Theta=\underset{\omega}{\max}\,G\left(\omega\right)$; r^{*} (или ρ^{*}) означает, что сканирование осу-

ществляется не сеткой параллельных плоскостей (или прямых), а одиночной плоскостью (или прямой) под разными углами наклона в пространстве (или плоскости). Для большей точности и надежности распознавания для каждого угла наклона сканирующих элементов можно производить сканирования 2–3 различными случайными параллельными плоскостями (прямыми), а полученные результаты усреднять. Например, $\mathrm{T}' = \sum_{i=1}^{3} \max_{t} (f(\theta, \rho_i, t))/3$ и $\mathrm{HyperT}' = \sum_{j=1}^{2} \sum_{i=1}^{2} \sum_{r} G(\omega_i, \varphi_j, r)/4$.

4. Максимальное количество пересечений исходного пространственного объекта сканирующей прямой:

$$\mathrm{Res}(F) = \mathrm{Hyper}\Theta \circ \mathrm{Hyper}\Omega \circ \mathrm{HyperT}\left(\Theta \circ \mathrm{T}\left(F_{\mathrm{sect}} \bigcap l(\theta, \rho)\right)\right),$$

где $\mathrm{T} = \underset{t}{\mathrm{rows}}(f(\theta, \rho = \rho^*, t))$; $\Theta = \underset{\theta}{\max} g(\theta)$; $\mathrm{HyperT} = G(\omega, \varphi, r = r^*)$; $\mathrm{Hyper}\Omega = \underset{\varphi}{\max} G(\omega, \varphi)$; $\mathrm{Hyper}\Theta = \underset{\omega}{\max} G(\omega)$; $\mathrm{rows}(f(X))$ — количество элементов дискретной функции $f(X)$ (количество пересечений прямой с 2D фигурой в плоскости сечения F_{sect}).

«Короткие» признаки 1–3 являются инвариантными к группе движений и сенситивными к масштабированию 3D изображения. «Короткий» признак 4 является инвариантным к группе движений и масштабированию 3D изображения. Стоит отметить, что «короткие» признаки 1–3 достаточно просто сделать инвариантными к масштабированию изображения при использовании функционалов в виде отношения функции и числа элементов в строке. Например, заменяя функционалы P и HyperP в «коротком» признаке 2 на функционалы $\mathrm{P} = (\mathrm{Row2D} + 1)\,\Delta\rho + \sum_{i=1}^{\mathrm{Row2D}-1} |g(\theta^*, \rho_{i+1}) - g(\theta^*, \rho_i)|/\mathrm{Row2D}$ и $\mathrm{HyperP} = \sum_{r} G(\omega^*, \varphi^*, r)\,\Delta r/\mathrm{Row3D}$, признак становится инвариантным к группе движения и масштабирования 3D изображения, где Row3D — количество ненулевых элементов в глубинной строке гипертрейс-матрицы 3TM.

Таким образом, «короткие» признаки имеют сокращенную форму композиции гипертриплетных и триплетных признаков для более быстрого их вычисления, так как сканирование большого числа объектов сеткой плоскостей со всех сторон и обработка сечений сеткой прямых со всех сторон достаточно емко по времени.

В заключение этого раздела подчеркнем разницу между «длинными» признаками верификации и «короткими» признаками фальсификации и их использованием. Признаки верификации способны описать любую информацию о пространственном объекте, а признаки фальсификации — только ограниченную часть информации. Признаки верификации описывают как индивидуальные, так и общие свойства 3D изображений, а признаки фальсификации — как правило, общие свойства, характерные и для других пространственных изображений данного класса. Признаки фальсификации вычисляются в десятки раз быстрее признаков верификации.

5 Результаты

Ввиду того, что статья носит только теоретический и концептуальный характер, описывает математическую модель и интеллектуальные возможности метода, реальные практические эксперименты и тестирование различных категорий признаков на различных базах 3D изображений с измерением времени ускорения вычисления в данной статье не проводились. Проверка свойств предложенного метода (различные вычислительные эксперименты) и практические результаты поиска 3D объектов в базе данных можно найти в [16, 22].

6 Заключение

Предлагаемый в настоящей статье новый геометрический метод сканирования и распознавания 3D изображений имеет множество способностей интеллектуального анализа и распознавания пространственных объектов. Так, конструируемые гипертриплетные признаки имеют композиционную структуру, которая способствует не только легкой машинной реализации этого алгоритма, но и конструированию большого числа признаков в автоматическом режиме. Данное обстоятельство особенно востребовано в зрительной системе робототехнике, когда машина должна самостоятельно анализировать и принимать решение [23].

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, описывающие метрические характеристики пространственного объекта). Возможность регулировать свойства построенных признаков заметно повышает интеллектуальные возможности гипертрейс-преобразования, что, несомненно, является его преимуществом [24].

Данный метод обладает определенной универсальностью, так как схема сканирования не привязана к геометрическим особенностям исходной пространственной модели. В связи с этим предлагаемая методика ориентирована на объекты любой сложности и конфигурации. Благодаря особенностям техники сканирования и аналитической структуры гипертриплетных признаков возможно конструирование признаков как инвариантных, так и сенситивных к группе движений и масштабным преобразованиям. Данное обстоятельство расширяет интеллектуальный анализ 3D изображений [24].

Одной из интеллектуальных способностей предлагаемого метода также является высокоуровневая предобработка, обработка и постобработка 3D изображения в одной технике сканирования, которая описана в [25].

Созданный математический инструмент для анализа 3D изображений — гипертрейс-матрица — позволяет распознавать пространственные 3D объекты сложной формы и структуры благодаря построению единой математической модели. В отличие от математического аппарата других методов данный инструмент позволяет параллельно с распознаванием объекта извлекать параметры его пространственной ориентации, положения и масштаба, не требуя для этого дополнительного сканирования [15].

В настоящей статье были описаны еще одни интеллектуальные способности гипертрейс-преобразования, а именно: принципы интеллектуального анализа и распознавания 3D изображений. Так, пространственный объект может быть очень быстро просканирован под одним углом наклона сетки плоскостей и одним углом наклона сетки прямых в плоскости сечения, по результатам такого сканирования вычисляется «короткий» признак 3D изображения. Если близость двух изображений по вычисленному вектору нескольких таких признаков не наблюдается, то исходное тестовое изображение нет смысла полностью сканировать, и оно исключается из дальнейшего рассмотрения. Программа автоматически начинает сканировать и распознавать следующий пространственный объект в зависимости от решаемой задачи.

Данное обстоятельство повышает скорость работы сканирующей системы и всей системы распознавания изображений в целом, так как большая часть кандидатов сразу исключается из рассмотрения. Кроме того, повышается надежность распознавания, ввиду того что заведомо разные кандидаты никогда не будут считаться похожими и не внесут искажения в усредненное изображение представителей своего класса, так как они уже исключены на ранних стадиях из рассмотрения.

Стоит отметить, что при использовании «коротких» признаков не производятся лишние сканирования 3D изображения, так как «длинные» и «короткие» признаки вычисляются в одной и той же технике сканирования (при вычислении «длинных» признаков дополнительно совершаются еще сканирования помимо тех, которые уже были использованы при вычислении «коротких» признаков).

Таким образом, использование гипертриплетных признаков разных категорий («коротких» и «длинных») заметно повышает интеллектуальные способности разрабатываемого метода распознавания 3D изображений, делая их на шаг ближе к человеческим.

Авторы планируют развить данный метод для 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 [26, 27]. Интеллектуальный уровень гипертрейс-преобразования может быть повышен благодаря развитию теории трейс-преобразования для интеллектуального анализа и распознавания деформированных и поврежденных 3D объектов, а также для распознавания двигающих изображений, когда одна часть изображения изменяет свое положение по отношению к другой части. Последнюю задачу не способен решить ни один из известных на сегодняшний момент методов.

Литература

[1] *Vasil'ev K. K., Dement'ev V. E., Andriyanov N. A.* Doubly stochastic models of images // Pattern Recogn. Image Anal. Adv. Math. Theor.Appl., 2015. Vol. 25. No. 1. P. 105–110.

[2] *Kiy K. I.* Segmentation and detection of contrast objects and their application in robot navigation // Pattern Recogn. Image Anal. Adv. Math. Theor. Appl., 2015. Vol. 25. No. 2. P. 338–346.

[3] *Myasnikov V. V.* Analysis of efficient linear local features of digital signals and images // Pattern Recogn. Image Anal. Adv. Math. Theor. Appl., 2016. Vol. 26. No. 1. P. 22–23.

[4] *Song S., Xiao J.* Sliding shapes for 3D object detection in depth images // Computer vision — ECCV 2014 / Eds. D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars. — Lecture notes in computer science ser. — Springer, 2014. Vol. 8694. P. 634–651.

[5] *Zhang Y., Song S., Tan P., Xiao J.* PanoContext: A whole-room 3D context model for panoramic scene understanding // Computer vision — ECCV 2014 / Eds. D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars. — Lecture notes in computer science ser. — Springer, 2014. Vol. 8694. P. 668–686.

[6] *Wang C., Huang K.-Q.* VFM: Visual feedback model for robust object recognition // J. Comput. Sci. Technol., 2015. Vol. 30. Iss. 2. P. 325–339.

[7] *Andreux M., Rodolà E., Aubry M., Cremers D.* Anisotropic Laplace–Beltrami operators for shape analysis // Computer vision — ECCV 2014 Workshop / Eds. L. Agapito, M. M. Bronstein, C. Rother. — Image processing, computer vision, patterns recognition, and graphics ser. — Springer, 2014. Vol. 8928. P. 299–312.

[8] *Lmaati E. A., Oirrak A. E., Kaddioui M. N., Ouahman A. A., Sadgal M.* 3D model retrieval based on 3D discrete cosine transform // Int. Arab J. Inform. Technol., 2010. Vol. 7. No. 3. P. 264–270.

[9] *Boucher M., Evans A. C., Siddiqi K.* Anisotropic diffusion of tensor fields for fold shape analysis on surfaces // Inform. Proc. Medical Imaging, 2011. Vol. 6801. P. 271–282.

[10] *Litman R., Bronstein A.* Learning spectral descriptors for deformable shape correspondence // Pattern Anal. Machine Intelligence, 2014. Vol. 36. Iss. 1. P. 171–180.

[11] *Elhachloufi M., Oirrak A. El., Aboutajdine D., Kaddioui M. N.* Affine invariant descriptors of 3D object using multiple regression model // Int. J. Comput. Sci. Inform. Technol., 2011. Vol. 3. Iss. 1. P. 1–10.

[12] *Баев А. А.* Методы распознавания 3D изображений на основе их кватернионных моделей. Дисс. ... канд. техн. наук. — Йошкар-Ола, 2011. 131 с.

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

[14] *Семов А. А.* Построение оптимальной стохастической равномерной сетки на сфере, инвариантной к повороту 3D изображения в пространстве // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XV Междунар. науч.-технич. конф. — Пенза: Изд-во АННОО «Приволжский дом знаний», 2015. С. 134–141.

[15] *Федотов Н. Г., Семов А. А.* Гипертрейс-матрица как основной инструмент анализа 3D-объектов // XXI век: итоги прошлого и проблемы настоящего плюс. Сер. Технические науки. Информационные технологии, 2015. Т. 1. No. 03(25). С. 63–69.

[16] *Fedotov N. G., Ryndina S. V., Semov A. A.* Trace transform of three-dimensional objects: Recognition, analysis and database search // Pattern Recogn. Image Anal. Adv. Math. Theor. Appl., 2014. Vol. 24. No. 4. P. 566–574.

[17] *Fedotov N. G.* The theory of image-recognition features based on stochastic geometry // Pattern Recogn. Image Anal. Adv. Math. Theor. Appl., 1998. Vol. 8. No. 2. P. 264–266.

[18] *Fedotov N., Romanov S., Goldueva D.* Application of triple features theory to the analysis of halftone images and colored textures. Feature construction along stochastic geometry and functional analysis // Comput. Inform. Sci., 2013. Vol. 6. No. 4. P. 17–24.

[19] *Федотов Н. Г., Семов А. А., Моисеев А. В.* 3D-трейс-преобразование: режимы сканирования, особенности стохастической реализации, способы ускорения вычислений // Известия высших учебных заведений. Поволжский регион. Технические науки, 2014. No. 3(31). С. 41–53.

[20] *Семов А. А.* Основные методы построения гипертрейс-матриц // XXI век: итоги прошлого и проблемы настоящего плюс, 2015. No. 3(25). С. 69–76.

[21] *Федотов Н. Г., Семов А. А., Моисеев А. В.* Минимизация признакового пространства распознавания 3D изображения на основе стохастической геометрии и функционального анализа // Машинное обучение и анализ данных, 2015. Т. 1. № 13. С. 1796–1814.

[22] *Семов А. А.* Экспериментальная проверка свойств 3D трейс-преобразования // XXI век: итоги прошлого и проблемы настоящего плюс, 2014. No. 3(19). С. 83–89.

[23] *Федотов Н. Г., Семов А. А.* Программный комплекс анализа и распознавания 3D изображений на основе пространственного трейс-преобразования со случайными параметрами сканирования. Свидетельство об официальной регистрации программ для ЭВМ № 2015612257 Роспатента от 16.02.15.

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

[25] *Федотов Н. Г., Семов А. А., Моисеев А. В.* Интеллектуальные возможности гипертрейс-преобразования: конструирование признаков с заданными свойствами // Машинное обучение и анализ данных, 2014. Т. 1. № 9. С. 1200–1214.

[26] *Fedotov N. G., Mokshanina D. A.* Recognition of halftone textures from the standpoint of stochastic geometry and functional analysis // Pattern Recogn. Image Anal. Adv. Math. Theor. Appl., 2010. Vol. 20. No. 4. P. 551–556.

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

Поступила в редакцию 19.07.2016

New method for three-dimensional images intelligent analysis and recognition: Description and examples

***N. G. Fedotov*[1], *A. A. Syemov*[2], *and A. V. Moiseev*[3]**

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

[1]Penza State University, 40 Krasnaya Str., Penza, Russia

[2]Comearth, 16 Gagarina Str., Penza, Russia

[3]Penza State Technological University, 1a Baidukova Proezd/11 Gagarina Str., Penza, Russia

Background: A new approach to the three-dimensional (3D) objects' recognition is proposed. A detailed mathematical description of method developed on the above approach basis is shown. Hypertrace transform technique scan is described and the scanning element choice is substantiated. The principles of 3D images intellectual analysis and recognition built on its basis are analyzed.

Methods: The suggested method is based on the theories elements of stochastic geometry and functional analysis. Hypertrace transform has many advantages and data mining capabilities. For example, one of the suggested method intellectual capabilities is the construction of different structure hypertriplet features ("long" and "short" features). Different types of features are reflected in the principles of 3D images intelligent analysis and recognition (verifiability and falsifiability of images).

Results: Due to only theoretical and conceptual article orientation, the practical results are missing. The theoretical examples description of verification of "long" features and falsification of "short" features of images is given. Their differences and practical application specificities are substantiated.

Concluding Remarks: Hypertrace transform has a unique ability which is a similar possibility of human visual system when at sufficiently brief glance, people quickly can distinguish two spatial objects from each other. This fact increases the scanning system speed and the image recognition system reliability in general, improving the intellectual abilities hypertrace transform.

Keywords: *hypertrace transform; 3D images intelligent analysis and recognition; invariant description; stochastic scan; hypertriplet feature analytical structure*

DOI: 10.21469/22233792.2.3.05

References

[1] Vasil'ev, K. K., V. E. Dement'ev, and N. A. Andriyanov. 2015. Doubly stochastic models of images. *Pattern Recogn. Image Anal. Adv. Math. Theor. Appl.* 25(1):105–110.

[2] Kiy, K. I. 2015. Segmentation and detection of contrast objects and their application in robot navigation. *Pattern Recogn. Image Anal. Adv. Math. Theor. Appl.* 25(2):338–346.

[3] Myasnikov, V. V. 2016. Analysis of efficient linear local features of digital signals and images. *Pattern Recogn. Image Anal. Adv. Math. Theor. Appl.* 26(1):22–23.

[4] Song S., Xiao J. 2014. Sliding shapes for 3D object detection in depth images *Computer vision — ECCV 2014.* Eds. D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Lecture notes in computer science ser. Springer. 8694:634–651.

[5] Zhang, Y., S. Song, P. Tan, and J. Xiao. 2014. PanoContext: A whole-room 3D context model for panoramic scene understanding. *Computer vision — ECCV 2014.* Eds. D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Lecture notes in computer science ser. Springer. 8694:668–686.

[6] Wang, C., and K.-Q. Huang. 2015. VFM: Visual feedback model for robust object recognition *J. Comput. Sci. Technol.* 30(2):325–339.

[7] Andreux, M., E. Rodolà, M. Aubry, and D. Cremers. 2014. Anisotropic Laplace–Beltrami operators for shape analysis. *Computer vision — ECCV 2014 Workshop*. Eds. L. Agapito, M. M. Bronstein, and C. Rother. Image processing, computer vision, pattern recognition, and graphics ser. Springer. 8928:299–312.

[8] Lmaati, E. A., A. E. Oirrak, M. N. Kaddioui, A. A. Ouahman, and M. Sadgal. 2010. 3D model retrieval based on 3D discrete cosine transform. *Int. Arab J. Inform. Technol.* 7(3):264–270.

[9] Boucher, M., A. C. Evans, and K. Siddiqi. 2011. Anisotropic diffusion of tensor fields for fold shape analysis on surfaces. *Inform. Proc. Medical Imaging* 6801:271–282.

[10] Litman, R., and A. Bronstein. 2014. Learning spectral descriptors for deformable shape correspondence. *Pattern Anal. Machine Intelligence* 36(1):171–180.

[11] Elhachloufi, M., A. El. Oirrak, D. Aboutajdine, and M. N. Kaddioui. 2011. Affine invariant descriptors of 3D object using multiple regression model. *Int. J. Comput. Sci. Inform. Technol.* 3(1):1–10.

[12] Baev, A. A. 2011. 3D images recognition methods based on its quaternion models. PhD Diss. Yoshkar-Ola. 131 p.

[13] Fedotov, N. G. 2009. *The theory of patterns recognition features based on stochastic geometry and functional analysis*. Moscow: Fizmatlit. 304 p.

[14] Syemov, A. A. 2015. Building optimal stochastic uniform grid on the sphere that are invariant to the 3D image rotation in space. *15th Scientific and Technical Conference (International) on Problems of Informatics in Education, Management, Economics and Technology Proceedings*. Penza: Privolzskiy Dom Znaniy Publs. 134–141.

[15] Fedotov, N. G., and A. A. Semov. 2015. Hypertrace-matrix as main tool for 3D objects analysis. *XXI Century: Past Results and Present Problems — Plus. Engineering Science. Information Technology* 03(25):63–69.

[16] Fedotov, N. G., S. V. Ryndina, and A. A. Semov. 2014. Trace transform of three-dimensional objects: Recognition, analysis and database search. *Pattern Recogn. Image Anal. Adv. Math. Theor. Appl.* 24(4):566–574.

[17] Fedotov, N. G. 1998. The theory of image-recognition features based on stochastic geometry. *Pattern Recogn. Image Anal. Adv. Math. Theor. Appl.* 8(2):264–266.

[18] Fedotov, N., S. Romanov, and D. Goldueva. 2013. Application of triple features theory to the analysis of half-tone images and colored textures. Feature construction along stochastic geometry and functional analysis. *Comput. Inform. Sci.* 6(4):17–24.

[19] Fedotov, N. G., A. A. Syemov, and A. V. Moiseev. 2014. 3D trace transform: Scan regimes, stochastic implementation particularities, accelerating calculations ways *Proc. Higher Educational Institutions. Volga region. Engineering science* 3(31):41–53.

[20] Semov, A. A. 2015. The basic methods of hypertrace-matrix formation. *XXI Century: Past Results and Present Problems — Plus. Engineering Science. Information Technology* 03(25):69–76.

[21] Fedotov, N. G., A. A. Syemov, and A. V. Moiseev. 2015. Feature space minimization of 3D image recognition based on stohastic geometry and functional analysis. *Machine Learning Data Anal.* 1(13):1796–1814.

[22] Semov, A. A. 2014. Experimental verification of 3D trace transform properties. *XXI Century: Past Results and Present Problems — Plus. Engineering Science. Information Technology* 03(19):83–89.

[23] Fedotov, N. G., and A. A. Syemov. February 16, 2015. Software for 3D images analysis and recognition based on the spatial trace transform with random scan parameters. Official registration certificate for computer programs No. 2015612257 of Rospatent.

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

[25] Fedotov, N. G., A. A. Syemov, and A. V. Moiseev. 2014. Intelligent capabilities hypertrace transform: Constructing features with predetermined properties. *Machine Learning Data Anal.* 1(9):1200–1214.

[26] Fedotov, N. G., and D. A. Mokshanina. 2010. Recognition of halftone textures from the standpoint of stochastic geometry and functional analysis. *Pattern Recogn. Image Anal. Adv. Math. Theor. Appl.* 20(4):551–556.

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

Received July 19, 2016

Бэггинг нейронных сетей в задаче анализа биологической активности ядерных рецепторов*

М. Р. Владимирова, М. С. Попова
mrvladimirova@gmail.com; popova@gmail.com
Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., д. 9

Работа посвящена решению проблемы повышения качества многозадачной классификации с помощью нейросетевой модели. Улучшение модели решения задачи проводится многозадачной моделью двухслойной нейронной сети. Рассматриваются две функции потерь: квадратичная и кросс-энтропийная. Для получения более точного результата в работе рассматривается композиция базовых классификаторов — бэггинг нейронных сетей. Сравнение моделей проводится с помощью вычислительного эксперимента на реальных данных, описывающих взаимодействия рецепторов и лиганд.

Ключевые слова: *клеточные рецепторы; биологическая активность; двухслойная нейронная сеть; бэггинг; многозадачность; разработка лекарств; кросс-энтропийная функция*

DOI: 10.21469/22233792.2.3.06

1 Введение

Рассматривается проблема многозадачной классификации на данных, описывающих взаимодействие ядерных рецепторов. Ядерные рецепторы представляют собой класс находящихся в клетках белков. Рецепторы влияют на транскрипцию генов: регулируют развитие, гомеостаз и обмен веществ в организме. Регулирование происходит в основном тогда, когда рецептор и лиганд — молекула, воздействующая на поведение рецептора, — взаимодействуют. Требуется предсказать, будет ли объект относиться к определенному классу, т. е. будет ли взаимодействовать данный лиганд с определенным рецептором. Проблема построения адекватных математических моделей для предсказания лиганд-рецепторного взаимодействия на основании данных о структурах химических соединений является актуальной задачей в фармакологии [1–4]. С помощью моделей проводится предварительная оценка характера взаимодействия лиганд и рецепторов, что позволяет снизить количество лабораторных экспериментов, необходимых для выявления активных лиганд.

Существуют два подхода к решению данной проблемы. Один из подходов заключается в компьютерном моделировании взаимодействия молекул, основанном на законах молекулярной динамики [5]. Такой способ является трудоемким и неприменим в случаях, когда точная трехмерная структура рецептора или лиганда неизвестна [6]. Второй подход — использование методов, основанных на статистике и машинном обучении. В литературе такой подход получил общее название «поиск количественных соотношений структура–свойство», или «Quantitative Structure–Activity Relationship» [7]. Модели, связывающие структуру лиганд с их биологической активностью, показали свою способность к предсказыванию лиганд-рецепторного взаимодействия [8, 9]. Точность модели машинного обучения зависит от размера обучающей выборки, поэтому для построения точных моделей необходим достаточный объем выборки. Несмотря на то что для некоторых рецепторов

*Проект поддержан грантом РФФИ № 16-07-01155.

уже проведено немало лабораторных экспериментов, данных о многих рецепторах оказывается недостаточно [10, 11]. Однако экспертные знания в области биохимии и фармакологии дают основания полагать, что факты связывания одних и тех же молекул с разными рецепторами не являются независимыми. Это означает, что можно компенсировать недостаток известных лиганд для данной цели наличием известных лиганд для подобных целей, используя многозадачное предсказание.

В данной работе решается набор взаимосвязанных или схожих задач обучения одновременно, с помощью алгоритмов обучения, имеющих схожее внутреннее представление, т. е. решается проблема многозадачной классификации. Информация о сходстве задач между собой позволяет совершенствовать алгоритм обучения и повышать качество решения основной задачи. Моделью классификации, позволяющей строить предсказания для группы рецепторов, предлагается использовать двухслойную нейронную сеть. Искусственные нейронные сети — эффективный инструмент решения исследовательских задач [8, 12–14]. Нейронные сети обладают уникальными особенностями, которые делают их надежными для решения задач с многомерными входными данными. Например, сети устойчивы к изменениям во входных данных [15], являются мультитасковыми, т. е. могут одновременно решать несколько задач [16], обучаются на всей выборке, не фрагментируя ее [17, 18].

Для повышения качества предсказаний лиганд-рецепторных взаимодействий предлагается использовать композицию двухслойных нейронных сетей. Одним из способов получения композиции классификаторов является использование бэггинга (bootstrap aggregating) [19]. Бэггинг генерирует из элементов обучающей выборки размера n семейство подвыборок размера n с помощью процедуры бутстрэп (bootstrap). Процедура основана на выборках с возвращениями, т. е. некоторые объекты могут встречаться в подвыборке более одного раза, а другие — отсутствовать. На каждой подвыборке настраивается классификатор. Ответы классификаторов агрегируются путем простого голосования. Бэггинг над базовыми алгоритмами позволяет увеличить точность и повысить устойчивость модели [20].

При решении задачи многоклассовой классификации на выходе нейронной сети необходимо получить вероятность принадлежности объекта каждому из классов. Рассмотрены две дифференцируемые функции потерь: квадратичная и кросс-энтропийная. Первая — сумма квадратов разности между истинным и восстановленным значениями. Чтобы знать суммарное число несовпадений между восстановленными метками классов и фактическими, используется кросс-энтропийная функция потерь — функция наибольшего правдоподобия в задаче логистической регрессии.

В работе был проведен вычислительный эксперимент на реальных данных, в ходе которого базовый алгоритм, двухслойная нейронная сеть, сравнивался с бэггингом над базовыми алгоритмами. Сравнение проводилось по значению функционала AUC (area under curve).

2 Постановка задачи

В задаче исследуется взаимодействие N лиганд с M рецепторами. Дана выборка $\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, \ldots, N\}$, состоящая из N пар объект–ответ.

Объектами $\mathbf{x}_i \in \mathbb{R}^K$ являются вектора признаковых описаний, в которых хранятся числовые свойства лиганд. Значения компонент вектора ответа $\mathbf{y}_i \in \{0, 1\}^M$ показывают, есть ли связь лиганда, соответствующего описанию $\mathbf{x}_i$, с различными рецепторами. Если реальный эксперимент не проводился или не дал адекватных результатов, то в ответе стоит пропуск. Назовем рецептор, взаимодействие с которым описывается m-м элементом

вектора ответа $\{y_i^m\}_{i=1}^N \in \{0,1\}$, m-рецептором, где $m \in \{1,\ldots,M\}$. Если лиганд с описанием $\mathbf{x}_i$ активирует m-рецептор, то $y_i^m = 1$, если не активирует — $y_i^m = 0$. Предположим, что $\mathbf{y}_i$ является реализацией случайного вектора, каждая компонента которого имеет распределение Бернулли. Исследуем взаимодействие каждого рецептора в разных задачах бинарной классификации. Пусть m-рецептору соответствует m-я задача, тогда решим одновременно M задач бинарной классификации, построив единую модель.

Базовые алгоритмы выбираются из класса двухслойных нейронных сетей:

$$\mathbf{z}(\mathbf{x},\boldsymbol{\theta}) = \mathbf{W}_2^{\mathsf{T}}\tanh(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}) : \mathbb{R}^K \to \mathbb{R}^H; \tag{1}$$

$$\mathbf{f}(\mathbf{x},\boldsymbol{\theta}) = \frac{1}{1+\mathbf{exp}\big(-\mathbf{z}(\mathbf{x},\boldsymbol{\theta})\big)} : \mathbb{R}^K \to [0,1]^M, \tag{2}$$

где $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W}_1^{\mathsf{T}}|\mathbf{W}_2^{\mathsf{T}})$ — вектор параметров двухслойной сети.

Значения признаков объекта $\mathbf{x}$ поступают на вход первому входному слою сети с весовой матрицей $\mathbf{W}_1$. Выходы первого слоя поступают на вход второму с весовой матрицей $\mathbf{W}_2$ — скрытому слою. Ответы на выходном слое интерпретируются как оценки вероятности того, что лиганд $\mathbf{x}$ связывается с рецепторами соответствующих задач:

$$\mathbf{f}(\mathbf{x},\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{P}(y_1 = 1|\mathbf{x},\boldsymbol{\theta}) \\ \mathsf{P}(y_2 = 1|\mathbf{x},\boldsymbol{\theta}) \\ \vdots \\ \mathsf{P}(y_M = 1|\mathbf{x},\boldsymbol{\theta}) \end{bmatrix}. \tag{3}$$

Выборка $\mathfrak{D}$ разделяется на две подвыборки — обучающую и контрольную. Для формирования бутстрэп-выборок $\mathfrak{L}_\ell$, $\ell = \{1,\ldots,L\}$, из обучающей выборки $\mathfrak{L}$ случайным образом отбирается несколько подмножеств, содержащих такое же количество элементов, как и исходное. Поскольку отбор производится случайно, набор элементов в этих выборках будет различным: некоторые из них могут быть отобраны по несколько раз, а другие — ни разу. Доля уникальных элементов в полученных выборках в среднем равна 0,56. На каждой из L выборок обучается базовый классификатор. Ответы классификаторов агрегируются путем простого голосования.

Моделью классификации $\mathbf{a}$, решающую одновременно M задач, назовем композицию базовых алгоритмов

$$\mathbf{a}(\mathbf{x},\boldsymbol{\theta},L) = \sum_{\ell=1}^{L} \pi_\ell \mathbf{f}(\mathbf{x},\boldsymbol{\theta}^\ell), \tag{4}$$

где $\pi_\ell = 1/L$ — веса базовых классификаторов; $\mathbf{f}(\mathbf{x},\boldsymbol{\theta}^\ell)$ — базовый алгоритм; $\boldsymbol{\theta}^\ell$ — вектор параметров базового алгоритма, вычисленного на подвыборке $\mathfrak{L}_\ell$.

Рассмотрим две задачи: линейную и логистическую регрессию. Каждой задаче соответствует функция ошибки, $\mathcal{L}_1$ и $\mathcal{L}_2$. Определим для каждой суммарную функцию потерь Q на некоторой подвыборке $\mathfrak{U}$ исходной выборки $\mathfrak{D}$ следующим образом:

$$\mathcal{L}_1(\boldsymbol{\theta},\mathbf{x}_i,\mathbf{y}_i) = \frac{1}{2}\sum_{m=1}^{M}\big(a^m(\mathbf{x}_i,\boldsymbol{\theta},L) - y_i^m\big)^2, \tag{5}$$

где $a^m(\mathbf{x}_i,\boldsymbol{\theta},L)$ — ответ классификатора на объекте $\mathbf{x}_i$ в m-й задаче, m-я компонента вектора $\mathbf{a}(\mathbf{x}_i,\boldsymbol{\theta},L)$ (4):

$$\mathcal{L}_2(\boldsymbol{\theta},\mathbf{x}_i,\mathbf{y}_i) = -\sum_{m=1}^{M} y_i^m \log \mathsf{P}(y_i^m = 1|\mathbf{x}_i,\boldsymbol{\theta}) + (1-y_i^m)\log \mathsf{P}(1-y_i^m = 1|\mathbf{x}_i,\boldsymbol{\theta}); \tag{6}$$

$$Q(\boldsymbol{\theta}, L|\mathfrak{U}) = \sum_{i=1}^{|\mathfrak{U}|} \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}_i, \mathbf{y}_i),\ \ \mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2\}.$$

Для нахождения оптимальных параметров $\hat{\mathbf{w}}$ и $\hat{L}$ модели $\mathbf{a}$ требуется решить задачу минимизации функции ошибки на обучающей выборке:

$$\hat{\boldsymbol{\theta}}, \hat{L} = \operatorname*{argmin}_{\boldsymbol{\theta}, L} Q(\boldsymbol{\theta}, L|\mathfrak{L}). \tag{7}$$

Для дополнительной оценки качества классификации будем вычислять значения функционала AUC на контрольной выборке для каждого класса по принципу один против всех и визуализировать полученные результаты с помощью ROC (receiver operating characteristic) кривых.

3 Оптимизация модели

Проанализируем построенную модель (1), (2), (4) с помощью декомпозиции ошибки Q на компоненты смещения и разброса (bias-variance decomposition) [21, 22]. Рассмотрим без потери общности декомпозицию функции ошибки для одной компоненты объекта выборки и одной задачи $a^m(x) = a(x)$ (4).

3.1 Квадратичная функция потерь

Пусть x — объект; y — истинная зависимость от объекта x; $f(x)$ — некоторый алгоритм, аппроксимирующий y. Квадратичной функции потерь (5) соответствует квадратичный риск

$$R(f) = \mathsf{E}_{x,y}\Big[\big(y - f(x)\big)^2\Big]. \tag{8}$$

Минимум среднеквадратичного риска достигается на функции, возвращающей условное матожидание ответа на фиксированном объекте. В случае бинарной классификации условие на минимум записывается следующим образом:

$$f^*(x) = \mathsf{E}[y|x] = \mathsf{P}(y = 1|x) = \operatorname*{argmin}_f R(f).$$

В работе рассматривается вероятностная модель (3). Вероятностная регрессионная модель лучше описывает предсказание вероятности биномиально распределенных величин в смысле среднеквадратичной ошибки, чем модель бинарной классификации:

$$\begin{aligned}\mathsf{E}\Big[\big(y - f(x)\big)^2\big|x\Big] &= \mathsf{E}\Big[\Big(\big(y - \mathsf{E}[y|x]\big) + \big(\mathsf{E}[y|x] - f(x)\big)\Big)^2\Big] = \\ &= \mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)^2\big|x\Big] + \Big(\mathsf{E}[y|x] - f(x)\Big)^2 + 2\,\mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)\big|x\Big] \cdot \Big(\mathsf{E}[y|x] - f(x)\Big) = \\ &= \mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)^2\big|x\Big] + \Big(\mathsf{E}[y|x] - f(x)\Big)^2 \geqslant \mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)^2\big|x\Big].\end{aligned}$$

Опишем зависимость среднеквадратичного риска (8) от выборки $\mathfrak{L}$ для композиции алгоритмов (4). Основной мерой качества алгоритма $a(x)$ возьмем усредненный по всем выборкам среднеквадратичный риск:

$$\mathcal{L}(a) = \mathsf{E}_{\mathfrak{L}}\Big[\mathsf{E}_{x,y}\Big[\big(y - a(x, \mathfrak{L})\big)^2\Big]\Big].$$

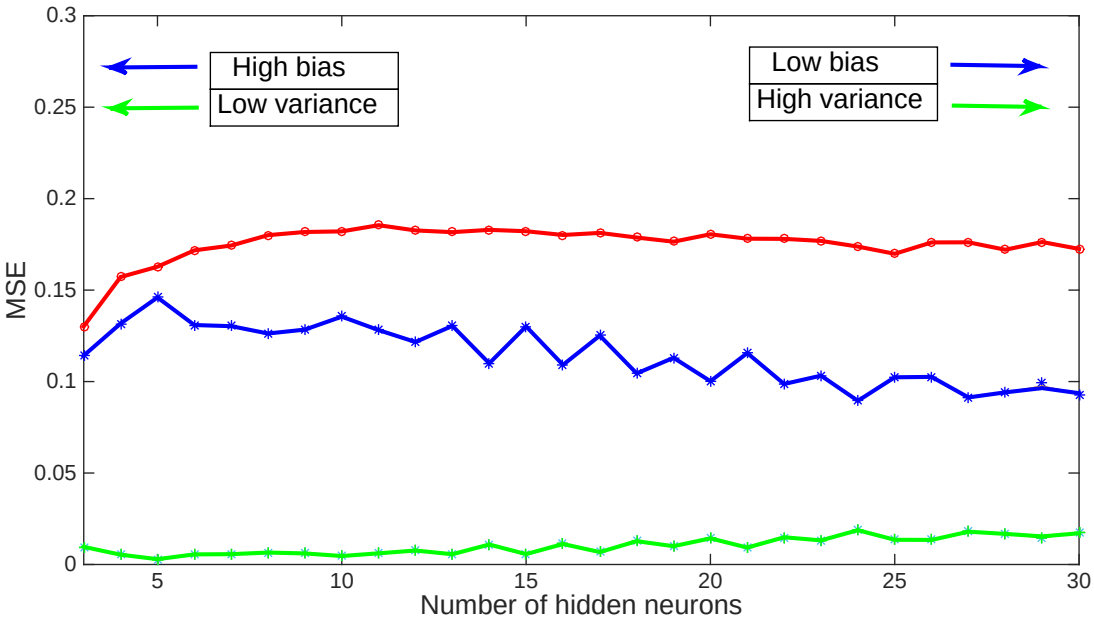


Рис. 1 Смещение и дисперсия базового алгоритма в зависимости от количества нейронов на скрытом слое

Для квадратичной функции ошибки для любого a $\mathcal{L}(a)$ представима в виде суммы из трех слагаемых [23]:

$$\mathcal{L}(a) = \mathsf{E}_{x,y}\Big[\big(y - \mathsf{E}\,[y|x]\big)^2\Big] + \mathsf{E}_{x,y}\Big[\Big(\mathsf{E}_{\mathfrak{L}}\big[a(x,\mathfrak{L})\big] - \mathsf{E}[y|x]\Big)^2\Big] + \\ + \mathsf{E}_{x,y}\Big[\mathsf{E}_{\mathfrak{L}}\Big[\Big(a(x,\mathfrak{L}) - \mathsf{E}_{\mathfrak{L}}\big[a(x,\mathfrak{L})\big]\Big)^2\Big]\Big]. \quad (9)$$

Первая компонента равна ошибке идеального алгоритма и описывает шум в данных. Невозможно построить алгоритм, имеющий меньшее ожидание ошибки. Вторая компонента характеризует смещение (bias) метода обучения, т. е. отклонение среднего ответа обученного алгоритма от ответа идеального алгоритма. Третья компонента характеризует дисперсию (variance), т. е. разброс ответов обученных алгоритмов относительно среднего ответа.

На рис. 1 показана визуализация зависимости смещения (синей линией) и дисперсии (зеленой линией) базового алгоритма от размерности пространства параметров, количества нейронов на скрытом слое. Также красной линией показана суммарная ошибка в зависимости от количества нейронов на скрытом слое. С увеличением количества скрытых нейронов смещение уменьшается, а дисперсия увеличивается.

Теорема 1. *Смещение композиции, полученной с помощью бэггинга, совпадает со смещением одного базового алгоритма* (2).

Доказательство.

$$\mathsf{E}_{x,y}\left[\left(\mathsf{E}_{\mathfrak{L}}\left[\frac{1}{L}\sum_{\ell=1}^{L} f(x,\mathfrak{L}_\ell)\right]-\mathsf{E}[y|x]\right)^2\right]=\mathsf{E}_{x,y}\left[\left(\frac{1}{L}\sum_{\ell=1}^{L}\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)-\mathsf{E}[y|x]\right]\right)^2\right]=$$
$$=\mathsf{E}_{x,y}\left[\left(\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)-\mathsf{E}[y|x]\right]\right)^2\right]=\mathsf{E}_{x,y}\left[\left(\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)\right]-\mathsf{E}[y|x]\right)^2\right]. \quad (10)$$

Таким образом, бэггинг не ухудшает смещенность модели. ■

Теорема 2. *Дисперсия композиции в L раз меньше дисперсии отдельных алгоритмов.*

Доказательство. Дисперсия композиции, построенной с помощью бэггинга, состоит из дисперсии одного базового алгоритма и корреляции между базовыми алгоритмами:

$$\mathsf{E}_{x,y}\left[\mathsf{E}_{\mathfrak{L}}\left[\left(\frac{1}{L}\sum_{\ell=1}^{L} f(x,\mathfrak{L}_\ell)-\mathsf{E}_{\mathfrak{L}}\left[\frac{1}{L}\sum_{\ell=1}^{L} f(x,\mathfrak{L}_\ell)\right]\right)^2\right]\right]=$$
$$=\frac{1}{L}\,\mathsf{E}_{x,y}\left[\mathsf{E}_{\mathfrak{L}}\left[\left(f(x,\mathfrak{L}_\ell)-\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)\right]\right)^2\right]\right]+$$
$$+\frac{L-1}{L}\,\mathsf{E}_{x,y}\left[\mathsf{E}_{\mathfrak{L}}\left[\left(f(x,\mathfrak{L}_\ell)-\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)\right]\right)\left(f(x,\mathfrak{L}_k)-\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_k)\right]\right)\right]\right]. \quad (11)$$

Если базовые алгоритмы некоррелированы, то дисперсия композиции в L раз меньше дисперсии отдельных алгоритмов. Поскольку нейронные сети относятся к неустойчивым моделям, корреляция алгоритмов отсутствует. ■

Таким образом, из теорем 1 и 2 следует, что бэггинг обеспечивает повышение точности.

3.2 Кросс-энтропийная функция потерь

Для случайных величин, имеющих распределение Бернулли, задается кросс-энтропийная функция потерь (6). Выразим данную функцию потерь через расстояние Кульбака–Лейблера. Рассмотрим задачу бинарной классификации $y=\{0,1\}$. Пусть p — истинная вероятность $\mathsf{P}(y=1|x)$ принадлежности объекта x к классу $y=1$; f — гипотетическая вероятность $\mathsf{P}(y=1|x)$, полученная с помощью алгоритма, аппроксимирующего y (3). Тогда расстояние Кульбака–Лейблера выражается следующим образом:

$$D_{\mathrm{KL}}(p,f)=p\ln\frac{p}{f}+(1-p)\ln\frac{1-p}{1-f}.$$

Обозначим $f^*(x)$ решение задачи

$$f^*(x)=\operatorname*{argmin}_{f\in[0,1]}\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\big(y,f\big)\right]. \quad (12)$$

Тогда получаем среднее геометрическое:

$$\ln\frac{f^*(x)}{1-f^*(x)}=\mathsf{E}_{x,y}\left[\ln\frac{f(x)}{1-f(x)}\right],$$

откуда

$$f^*(x)=\frac{1}{Z}\exp\left(\mathsf{E}_{x,y}\left[\ln f(x)\right]\right),$$

где Z — нормировочная константа, не зависящая от y.

Основной мерой качества алгоритма $a(x)$ возьмем усредненное по всем выборкам расстояние Кульбака–Лейблера:

$$\mathcal{L}(a) = \mathsf{E}_{\mathfrak{L}}\left[\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\left(y, a(x, \mathfrak{L})\right)\right]\right]. \tag{13}$$

Для решения задачи (12) необходимо разложить (13) на шум, смещение и дисперсию, как это было сделано для квадратичной функции потерь (9).

Теорема 3. *Ошибкой идеального алгоритма является энтропия от истинной вероятности $H(p)$.*

Доказательство. Расстояние между истинными ответами y и истинной вероятностью принадлежности объекта к классам p будет являться шумом в данных (ошибкой идеального алгоритма):

$$\begin{aligned}\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}(y,p)\right] &= \mathsf{E}_{x,y}\left[y\ln\frac{y}{p} + (1-y)\ln\frac{1-y}{1-p}\right] = \\ &= \mathsf{E}_{x,y}\left[y\ln y - y\ln p + (1-y)\ln(1-y) - (1-y)\ln(1-p)\right] = \\ &= -p\ln p - (1-p)\ln(1-p) = H(p),\end{aligned}$$

где $H(p)$ — функция энтропии. ■

Утверждение 1. *Смещение B и дисперсия V для функции ошибки $\mathcal{L}$ выражаются следующим образом* [22]*:*

$$\begin{aligned} B &= \mathcal{L}\left(p, a^*(x)\right), & a^*(x) &= \operatorname*{argmin}_{a\in[0,1]} \mathcal{L}(y,a); \\ V &= \mathsf{E}_{x,y}\left[\mathcal{L}\left(a^+(x), a^*(x)\right)\right], & a^+(x) &= \operatorname*{argmin}_{a\in[0,1]} \mathcal{L}(p,a). \end{aligned}$$

Из теоремы 3 и утверждения 1 получаем, что выражение (13) представляется в виде суммы трех слагаемых:

$$\mathcal{L}(a) = \mathsf{E}_{\mathfrak{L}}\left[\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\left(y, a(x,\mathfrak{L})\right)\right]\right] = H(p) + D_{\mathrm{KL}}\left(p, a^*(x)\right) + \mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\left(\mathsf{E}_{\mathfrak{L}}\left[a(x,\mathfrak{L})\right], a^*(x)\right)\right].$$

Тогда смещение бэггинга совпадает со смещением одного базового алгоритма, а дисперсия бэггинга уменьшается [22].

Таким образом, проводя сравнение разложений ошибки между композицией алгоритмов и одним базовым алгоритмом, получили, что для обеих функций потерь $\mathcal{L}_1$ и $\mathcal{L}_2$ выполняется равенство смещений и уменьшение дисперсий. Это означает, что результаты бэггинга нейронных сетей должны быть точнее, чем отдельной нейронной сети. Подтвердим вышеизложенные теорические выкладки вычислительным экспериментом.

3.3 Нахождение параметров модели

Оптимизация вектора параметров $\boldsymbol{\theta}$, минимизирующего суммарную функцию потерь (7) по обучающей выборке $\mathfrak{L}$, проводится модифицированным методом обратного распространения ошибки. Псевдокод алгоритма в представлен в Алгоритме 1.

Алгоритм 1 Модифицированный метод обратного распространения ошибки

Вход: выборка $\mathfrak{D} = \left\{(\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N\right\}$, количество циклов C, число нейронов в скрытом слое H, темп обучения сети η, модель с заданными функциями активации на первом α_1 и втором слоях α_2;

Выход: весовые параметры w_{jh}, w_{hm};

инициализировать веса w_{jh}, w_{hm};
задать $k = 0$;
повторять
выбрать объект $\mathbf{x}_i$ из $\mathfrak{D}$;
прямой ход:
вычислить значение функции на скрытом слое $u_i^h := \alpha_{1h}\left(\sum_{j=0}^n w_{jh}x_i^j\right),\ h = 1, \ldots, H$,
вычислить значение функции на выходном слое $f_i^m := \alpha_{2m}\left(\sum_{h=0}^H w_{hm}u_i^h\right),\ m = 1, \ldots, M$,
если есть результаты экспериментов: $y_i^m = 0$ или $y_i^m = 1$ **то**
вычислить значение ошибки на выходном слое ε_i^m,
функция ошибки квадратичная: $\varepsilon_i^m := f_i^m - y_i^m$,
функция ошибки кросс-энтропийная: $\varepsilon_i^m := -y_i^m / f_i^m - (1 - y_i^m) / (1 - f_i^m)$;
иначе
обработка пропусков $\varepsilon_i^m := 0$,
обратный ход:
вычислить значение ошибки на скрытом слое $\varepsilon_i^h := \sum_{m=1}^M \varepsilon_i^m \alpha'_{2m} w_{hm},\ h = 1, \ldots, H$,
α'_2 — функция, обратная к функции активации;
градиентный шаг:
$w_{hm} := w_{hm} - \eta\varepsilon_i^m \alpha'_{2m} u_i^h,\ h = 0, \ldots H,\ m = 1, \ldots, M$,
$w_{jh} := w_{jh} - \eta\varepsilon_i^m \alpha'_{1h} x_i^j,\ j = 0, \ldots n,\ h = 1, \ldots, H$;
$k := k + 1$;
пока $k < C$;

4 Вычислительный эксперимент

Выборка $\mathfrak{D}$ состоит из описания взаимодействия $N = 8513$ лиганд с $M = 12$ рецепторами: NR-AhR, NR-AR-LBD, NR-AR, SR-MMP, NR-ER, SR-HSE, SR-p53, NR-PPAR-gamma, SR-ARE, NR-Aromatase, SR-ATAD5 и NR-ER-LBD. Каждый объект описан $K = 185$ признаками. Биологическая активность выражается бинарным значением ответов: 1 — есть взаимодействие; 0 — нет взаимодействия. Если реальный эксперимент не проводился или не дал результатов, то в ответе стоит пропуск. На рис. 2 указано распределение объектов по классам. Около половины объектов — с известым бинарным ответом. Доля полностью размеченных объектов составляет 16% исходной выборки. Объект с пропуском в ответе не участвовал в тестировании.

Проведен вычислительный эксперимент на реальных данных, представленных в выборке $\mathfrak{D}$. Цель эксперимента — проверить адекватность работы базового алгоритма; получить оценки качества, необходимые для сравнения с предложенной в работе моделью классификации; сравнить качество результатов, полученных с помощью базового алгоритма и предложенной модели классификации.

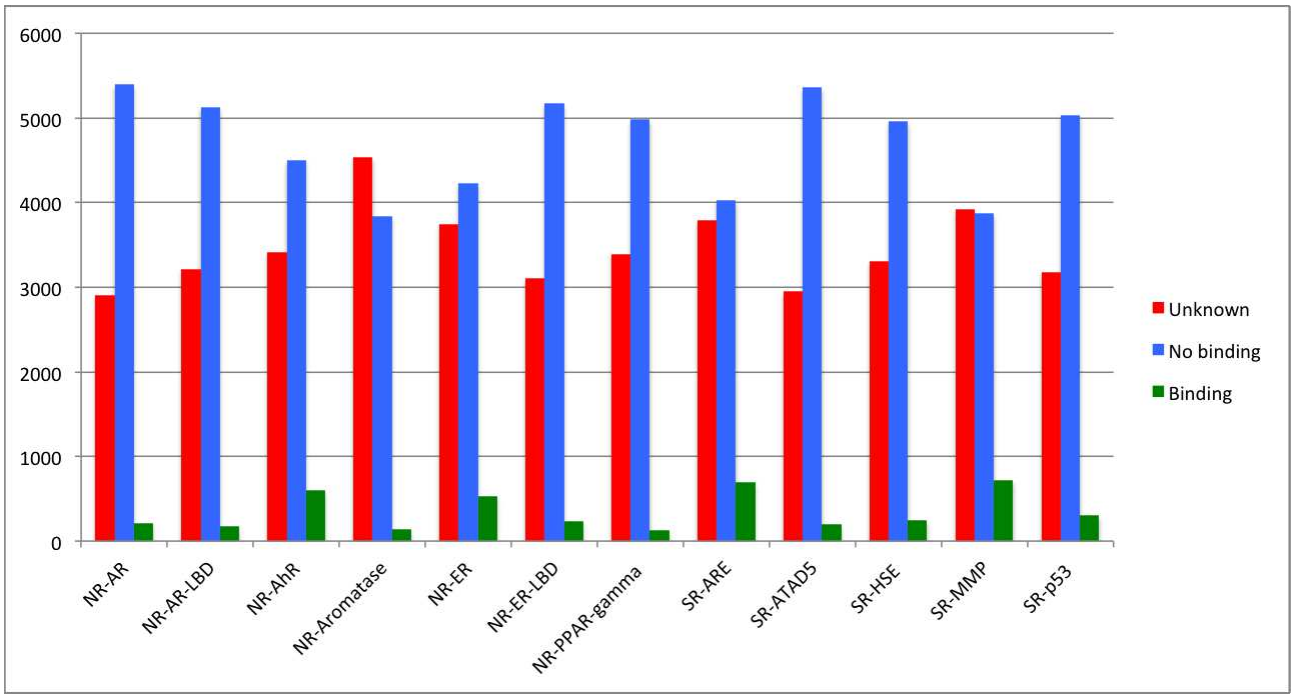


Рис. 2 Количество связывающихся лиганд для каждого рецептора

4.1 Базовый эксперимент

Для оценки качества результата была использована кросс-валидация обучающей выборки на 5 непересекающихся блоков. Для проверки качества алгоритма использованы ROC-кривые. На контрольных выборках вычислены значения функционала AUC. Для каждого рецептора синий линией на рис. 3 изображена ROC-кривая с вычисленным значением AUC.

Таким образом, нейронная сеть показала свою способность предсказывать биологическую активность лиганд и рецепторов.

4.2 Настройка параметров

Для улучшения качества классификации нейронной сети проведена настройка параметра H числа нейронов на скрытом слое. Значение функционала AUC вычислялось в зависимости от числа нейронов из промежутка от 1 до 100 с шагом, равным 5. На рис. 4 приведены зависимости для первых трех рецепторов. Результаты на большинстве рецепторов незначительно меняются в зависимости от H, но на некоторых точность классификации увеличивается, как для рецептора NR-AhR. Возьмем значение $H = 100$, при котором значение функционала AUC стабилизируется и остается примерно константой.

Параметр $\hat{L}$ находится с помощью анализа графика зависимости значения функционала AUC от количества разбиений в модели классификации (см. (7)). На графике для рецептора NR-AhR на рис. 5 видно, что с увеличением числа разбиений значение AUC растет, но с определенного момента значение остается константным. Проанализировав графики для всех 12 рецепторов, выбираем $L = 100$, при котором для каждого рецептора значение AUC на графике становится постоянным.

4.3 Бэггинг

Проведен вычислительный эксперимент для предложенной модели бэггинга с оптимальными параметрами. Результаты эксперимента показаны на рис. 3. Красной линией

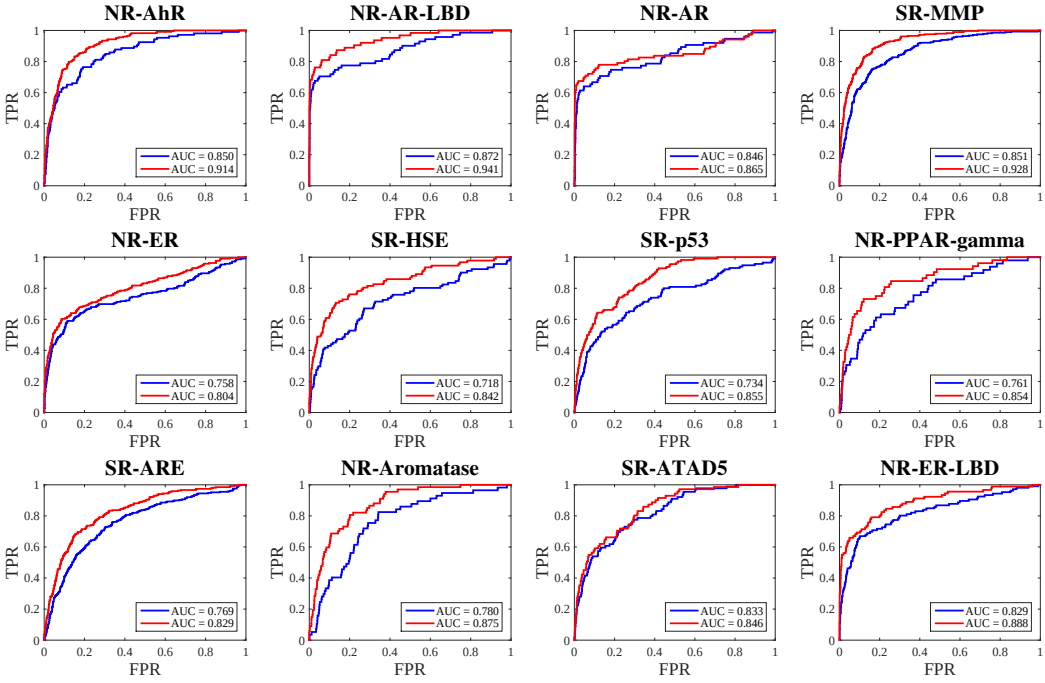


Рис. 3 ROC-кривые базового алгоритма и бэггинга

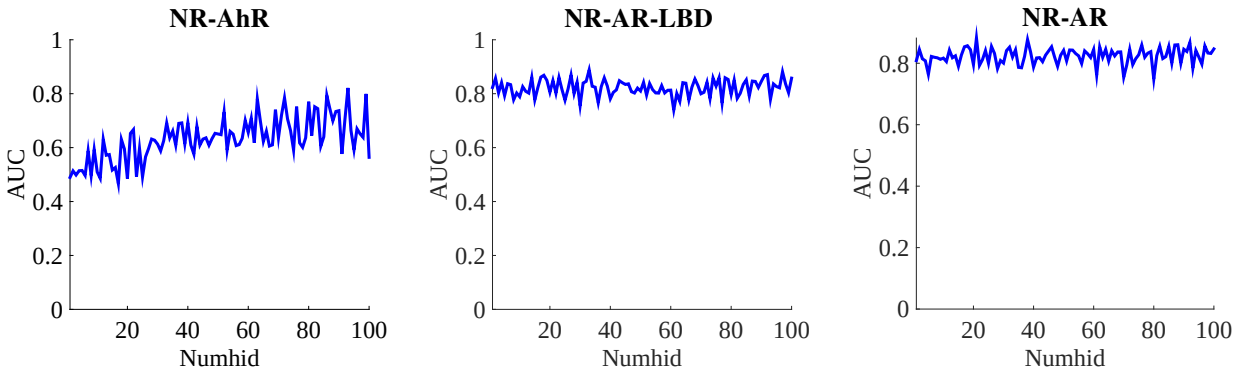


Рис. 4 Графики зависимости значения функционала AUC от количества нейронов на скрытом слое

изображены ROC-кривые бэггинга с вычисленным значением AUC. Площадь под кривой базового алгоритма для каждого рецептора меньше площади под соответствующей кривой бэггинга. Сравнение значений функционала AUC, полученного с помощью базового алгоритма нейронной сети и предложенного алгоритма бэггинга нейронных сетей, приведено в табл. 1. Сравнение проведено также между двумя функциями потерь: кросс-энтропийной и квадратичной.

Таким образом, из итоговых графиков на рис. 3 и табл. 1 видно, что предложенный алгоритм повысил качество классификации. Для рецепторов SR-HSE, SR-p53, NR-PPAR-gamma и NR-Aromatase качество увеличилось на 8%–12%, для NR-AhR, NR-AR-LBD, SR-MMP, NR-ER, SR-ARE и NR-ER-LBD — на 4%–7%, для SR-AR-LBD и SR-ATAD5 —

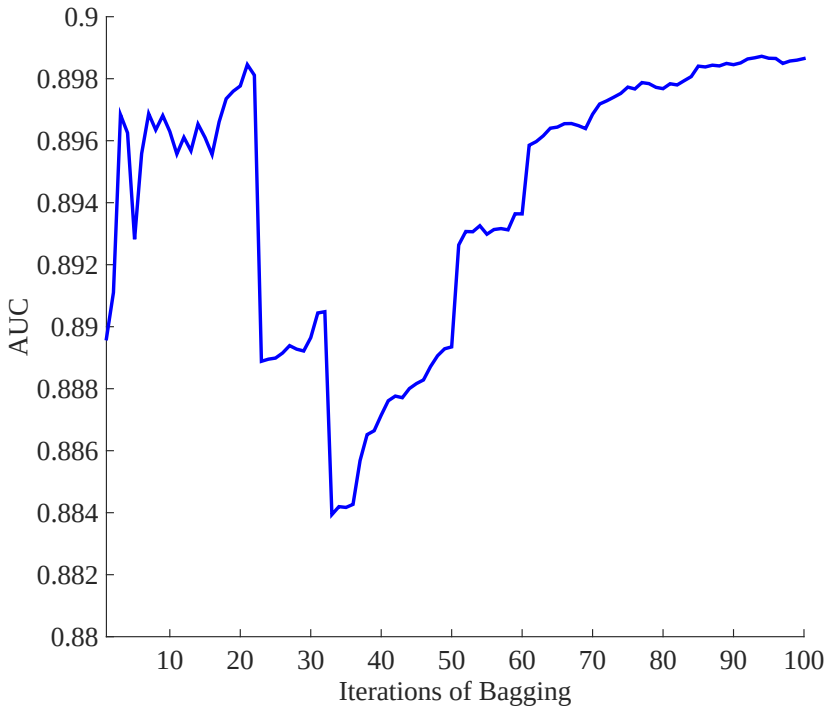


Рис. 5 График зависимости значения функционала AUC от количества разбиений на выборки в бэггинге

Таблица 1 Сравнение значений функционала AUC базового алгоритма нейронной сети и предложенного алгоритма бэггинга нейронных сетей с двумя функциями потерь: кросс-энтропийной и квадратичной

Рецептор	Нейронная сеть (кросс-энтропия)	Бэггинг (кросс-энтропия)	Нейронная сеть (квадратичная)	Бэггинг (квадратичная)
NR-AhR	0,8589 ± 0,0216	**0,9089 ± 0,0210**	0,8584 ± 0,0150	**0,9088 ± 0,0174**
NR-AR-LBD	0,8725 ± 0,0455	**0,9138 ± 0,0064**	0,9008 ± 0,0490	**0,9207 ± 0,0458**
NR-AR	0,8456 ± 0,0294	**0,8658 ± 0,0129**	0,8457 ± 0,0312	**0,8704 ± 0,0166**
SR-MMP	0,8512 ± 0,0483	**0,9132 ± 0,0110**	0,8651 ± 0,0080	**0,9161 ± 0,0109**
NR-ER	0,7585 ± 0,0726	**0,8109 ± 0,0329**	0,7545 ± 0,0414	**0,8151 ± 0,0253**
SR-HSE	0,7189 ± 0,0583	**0,8274 ± 0,0193**	0,7541 ± 0,0176	**0,8380 ± 0,0347**
SR-p53	0,7345 ± 0,0838	**0,8532 ± 0,0257**	0,7660 ± 0,0236	**0,8585 ± 0,0204**
NR-PPAR-gamma	0,7610 ± 0,0725	**0,8435 ± 0,0437**	0,7818 ± 0,0285	**0,8539 ± 0,0171**
SR-ARE	0,7698 ± 0,0307	**0,8265 ± 0,0208**	0,7652 ± 00309	**0,8268 ± 0,0076**
NR-Aromatase	0,7808 ± 0,0482	**0,8697 ± 0,0308**	0,8466 ± 0,0531	**0,8676 ± 0,0218**
SR-ATAD5	0,8338 ± 0,0714	**0,8682 ± 0,0187**	0,7713 ± 0,0648	**0,8629 ± 0,0332**
NR-ER-LBD	0,8299 ± 0,0241	**0,8917 ± 0,0267**	0,8515 ± 0,0251	**0,8884 ± 0,0168**

на 2% и 3% соответственно. Сравнивая результаты, полученные с разными функциями потерь, получаем, что значения AUC для одних и тех же рецепторов различаются максимум на 1,1% у рецептора SR-HSE, что меньше средней погрешности результатов.

Проведен вычислительный эксперимент для бэггинга, мощности подвыборок которого меньше мощности исходной выборки. Точность результатов падает с уменьшением размера подвыборки.

5 Заключение

В работе решалась проблема предсказания лиганд-рецепторного взаимодействия. В качестве модели классификации была предложена композиция двухслойных нейронных сетей — бэггинг. Рассмотрены задачи линейной и логистической регрессии с квадратичной и кросс-энтропийной функциями потерь соответственно. Исследовано изменение качества классификации с помощью декомпозиции функции ошибки на смещение и дисперсию. Проведено сравнение моделей с помощью вычислительного эксперимента на реальных данных. Полученные результаты говорят о том, что бэггинг позволяет повысить качество классификации.

Авторы выражают благодарность В. В. Стрижову за постановку задачи и внимательное отношение к работе.

Литература

[1] *Perkins R., Fang H., Tong W., Welsh W. J.* Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology // Environ. Toxicol. Chem., 2003. Vol. 22. No. 8. P. 1666–1679.

[2] *Bhasin M., Raghava G. P. S.* Classification of nuclear receptors based on amino acid composition and dipeptide composition // J. Biol. Chem., 2004. Vol. 279. No. 22. P. 23262–23266.

[3] *Salum L. B., Andricopulo A. D.* Fragment-based QSAR: Perspectives in drug design // Mol. Divers., 2009. Vol. 13. No. 3. P. 277–285.

[4] *Myint K. Z., Xie X. Q.* Recent advances in fragment-based QSAR and multi-dimensional QSAR methods // Int. J. Mol. Sci., 2010. Vol. 11. No. 10. P. 3846–3866.

[5] *Brown R. D., Martin Y. C.* The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding // J. Chem. Inf. Comp. Sci., 1997. Vol. 37. No. 1. P. 1–9.

[6] *DiMasi J. A., Hansen R. W., Grabowski H. G.* The price of innovation: New estimates of drug development costs // J. Health Econ., 2003. Vol. 22. No. 2. P. 151–185.

[7] *Zhang L., Zhu H., Oprea T. I., Golbraikh A., Tropsha A.* QSAR modeling of the blood-brain barrier permeability for diverse organic compounds // Pharm. Res., 2008. Vol. 25. No. 8. P. 1902–1914.

[8] *Myint K. Z., Wang L., Tong Q., Xie X. Q.* Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions // Mol. Pharm., 2012. Vol. 9. No. 10. P. 2912–2923.

[9] *Zhang L., Fourches D., Sedykh A., et al.* Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening // J. Chem. Inf. Model., 2013. Vol. 53. No. 2. P. 475–492.

[10] *Enrique Sucar L., Bielza C., Morales E. F., Hernandez-Leal P., Zaragoza J. H., Larrañaga P.* Multi-label classification with Bayesian network-based chain classifiers // Pattern Recogn. Lett., 2014. Vol. 41. No. 1. P. 14–22.

[11] *Popova M.* Feature selection and multi-task prediction of biological activity for nuclear receptors, technical report. URL: `https://goo.gl/5nXQMZ`.

[12] *Barkoula N. M., Alcock B., Cabrera N. O., Peijs T.* Fatigue properties of highly oriented polypropylene tapes and all-polypropylene composites // Polym. Polym. Compos., 2008. Vol. 16. No. 2. P. 101–113.

[13] *Steffen C., Thomas K., Huniar U., Hellweg A., Rubner O., Schroer A.* TmoleX — a graphical user interface for Turbomole. // J. Comput. Chem., 2010. Vol. 31. No. 16. P. 2967–2970.

[14] *Fang J., Yang R., Gao L., et al.* Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery // Mol. Divers., 2015. Vol. 19. No. 1. P. 149–162.

[15] *Gonzalez-Diaz H., Bonet I., Teran C., et al.* ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds // Eur. J. Med. Chem., 2007. Vol. 42. No. 5. P. 580–585.

[16] *Patra J. C., Chua K. H. K.* Neural network based drug design for diabetes mellitus using QSAR with 2D and 3D descriptors // Joint Conference (International) on Neural Networks Proceedings, 2010. P. 18–23.

[17] *Tu J. V.* Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes // J. Clin. Epidemiol., 1996. Vol. 49. No. 11. P. 1225–1231.

[18] *Lisboa P.* A review of evidence of health benefit from artificial neural networks in medical intervention // Neural Networks, 2002. Vol. 15. No. 1. P. 11–39.

[19] *Ha K., Cho S., Maclachlan D.* Response models based on bagging neural networks // J. Interact. Mark., 2005. Vol. 19. No. 1. P. 17–30.

[20] *Zhou Z. H., Wu J., Tang W.* Ensembling neural networks: Many could be better than all // Artif. Intell., 2002. Vol. 137. No. 1-2. P. 239–263.

[21] *Tibshirani R.* Bias, variance and prediction error for classification rules. University of Toronto, Department of Statistics, 1996.

[22] *James G. M.* Variance and bias for general loss functions // Mach. Learn., 2001. Vol. 51. No. 2. P. 115–135.

[23] *Geman S., Bienenstock E., Doursat R.* Neural networks and the bias/variance dilemma. — 1992. P. 1–58.

Поступила в редакцию 15.09.2016

Bagging of neural networks for analysis of nuclear receptor biological activity*

M. R. Vladimirova and M. S. Popova

mrvladimirova@gmail.com; popova@gmail.com

Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, Russia

The paper is devoted to the multitask classification problem. The main purpose is building an adequate model to predict whether the object belongs to a particular class, precisely, whether the ligand binds to a specific nuclear receptor. Nuclear receptors are a class of proteins found within cells. These receptors work with other proteins to regulate the expression of specific genes, thereby controlling the development, homeostasis, and metabolism of the organism. The regulation of gene expression generally only happens when a ligand — a molecule that effects the receptor's behavior — binds to a nuclear receptor. Two-layer neural network is used as a classification model. The paper considers the problems of linear and logistic regressions with squared and cross-entropy loss functions. To analyze the classification result, the authors propose to decompose the error into bias and variance terms. To improve the quality of classification by reducing the error variance, they suggest the composition of neural networks: the bagging procedure. The proposed method improves the quality of the investigated sample classification.

Keywords: *nuclear receptors; biological activity; two-layer neural network; bagging; multitask learning; drug-design; cross-entropy*

DOI: 10.21469/22233792.2.3.06

*This research is funded by the Russian Foundation for Basic Research, grant 16-07-01155.

References

[1] Perkins, R., H. Fang, W. Tong, and W. J. Welsh. 2003. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* 22(8):1666–1679.

[2] Bhasin, M., and G. P. S. Raghava. 2004. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279(22):23262–23266.

[3] Salum, L. B., and A. D. Andricopulo. 2009. Fragment-based QSAR: Perspectives in drug design. *Mol. Divers.* 13(3):277–285.

[4] Myint, K. Z., and X. Q. Xie. 2010. Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *Int. J. Mol. Sci.* 11(10):3846–3866.

[5] Brown, R. D., and Y. C. Martin. 1997. The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comp. Sci.* 37(1):1–9.

[6] DiMasi, J. A., R. W. Hansen, and H. G. Grabowski. 2003. The price of innovation: New estimates of drug development costs. *J. Health Econ.* 22(2):151–185.

[7] Zhang, L., H. Zhu, T. I. Oprea, A. Golbraikh, and A. Tropsha. 2008. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* 25(8):1902–1914.

[8] Myint, K. Z., L. Wang, Q. Tong, and X. Q. Xie. 2012. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol. Pharm.* 9(10):2912–2923.

[9] Zhang, L., D. Fourches, A. Sedykh, *et al.* 2013. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J. Chem. Inf. Model.* 53(2):475–492.

[10] Enrique Sucar, L., C. Bielza, E. F. Morales, P. Hernandez-Leal, J. H. Zaragoza, and P. Larrañaga. 2014. Multi-label classification with Bayesian network-based chain classifiers. *Pattern Recogn. Lett.* 41(1):14–22.

[11] Popova, M. *Feature selection and multi-task prediction of biological activity for nuclear receptors, technical report.* Available at: `https://goo.gl/5nXQMZ` (accessed December 20, 2015).

[12] Barkoula, N. M., B. Alcock, N. O. Cabrera, and T. Peijs. 2008. Fatigue properties of highly oriented polypropylene tapes and all-polypropylene composites. *Polym. Polym. Compos.* 16(2):101–113.

[13] Steffen, C., K. Thomas, U. Huniar, A. Hellweg, O. Rubner, and A. Schroer. 2010. TmoleX — a graphical user interface for Turbomole. *J. Comput. Chem.* 31(16):2967–2970.

[14] Fang, J., R. Yang, L. Gao, *et al.* 2015. Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery. *Mol. Divers.* 19(1):149–162.

[15] Gonzalez-Diaz, H., I. Bonet, C. Teran, *et al.* 2007. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* 42(5):580–585.

[16] Patra, J. C., and K. H. K. Chua. 2010. Neural network based drug design for diabetes mellitus using QSAR with 2D and 3D descriptors. *Joint Conference (International) on Neural Networks Proceedings.* 18–23.

[17] Tu, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* 49(11):1225–1231.

[18] Lisboa, P. 2002. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks* 15(1):11–39.

[19] Ha, K., S. Cho, and D. Maclachlan. 2005. Response models based on bagging neural networks. *J. Interact. Mark.* 19(1):17–30.

[20] Zhou, Z. H., J. Wu, and W. Tang. 2002. Ensembling neural networks: Many could be better than all. *Artif. Intell.* 137(1-2):239–263.

[21] Tibshirani, R. 1996. Bias, variance and prediction error for classification rules. University of Toronto, Department of Statistics.

[22] James, G. M. 2001. Variance and bias for general loss functions. *Mach. Learn.* 51(2):115–135.

[23] Geman, S., E. Bienenstock, and R. Doursat. 1992. Neural networks and the bias/variance dilemma. 1–58.

Received September 15, 2016