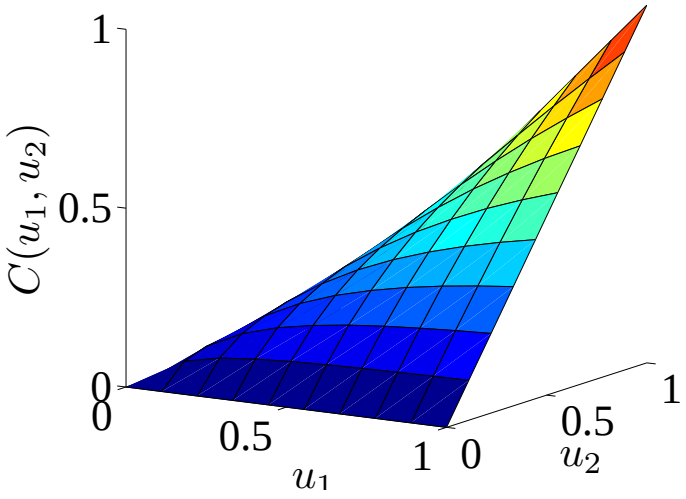
ISSN 2223-3792

Машинное обучение и анализ данных

Декабрь 2012 **Том 1, номер 4**



Машинное обучение и анализ данных

Journal of Machine Learning and Data Analysis

ISSN 2223-3792 Rus

Цель журнала — развитие теории машинного обучения и интеллектуального анализа данных и методов проведения вычислительных экспериментов. Журнал публикует новые теоретические и обзорные статьи с результатами научных исследований в области информатики и приложений. Принимаются статьи на русском и английском языке. Журнал включен в российский индекс научного цитирования РИНЦ.

Тематика журнала:
— регрессионный анализ,
— классификация,
— кластеризация,
— многомерный статистический анализ,
— байесовские методы регрессии и классификации,
— методы прогнозирования временных рядов,
— методы оптимизации в задачах машинного обучения и анализа данных,
— методы визуализации данных,
— обработка и распознавание речи и изображений,
— анализ и понимание текста, информационный поиск,
— прикладные задачи анализа данных.

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

Москва, 2012

Содержание

Параметрическое семейство базовых скелетов многоугольной фигуры*

Жукова К. В., Рейер И. А.
kz@pisem.net, reyer@forecsys.ru
Москва, Вычислительный центр РАН

В работе рассматривается базовый скелет — устойчивое скелетное представление формы, строящееся на основе скелета аппроксимирующей объект многоугольной фигуры. Исследуются свойства монотонности и непрерывности изменения базового скелета при увеличении величины точности аппроксимации. Вводится понятие разметки скелета — множества точек скелета многоугольной фигуры, описывающего процесс изменения базового скелета и позволяющего строить базовые скелеты для заданного набора или интервала значений точности аппроксимации.

Ключевые слова: скелетное представление формы, регуляризация скелета, базовый скелет, размеченный скелет, масштабируемые модели формы.

Parametric family of skeleton bases of a polygonal figure*

Zhukova K. V., Reyer I. A.
Computing Center of the Russian Academy of Sciences

In the paper, a skeleton base is considered. A skeleton base is a stable skeletal shape representation constructed with use of a polygonal figure approximating the shape. The monotonicity and continuity of change of a skeleton base with growth of the approximation accuracy value is investigated. A concept of a skeleton markup is presented. A skeleton markup is a set of points of a polygonal figure's skeleton describing the change of a skeleton base and allowing one to build skeleton bases for a given set or range of approximation accuracy values.

Keywords: *skeletal shape representation, skeleton regularization, skeleton base, marked skeleton, scalable shape models.*

Введение

Скелетное представление формы объекта [1] является широко используемым инструментом при решении задач распознавания изображений, поскольку скелет содержит информацию о структуре объекта. Однако использовать такое представление в «чистом» виде не представляется возможным, так как оно очень чувствительно к изменениям границы. Даже при небольших «шевелениях» границы в ее окрестности возникают шумовые скелетные ветви, в результате чего разница между скелетами близких объектов может быть очень существенной. При этом скелет содержит некоторую «фундаментальную» часть, на которую изменения границы в определенных пределах влияют незначительно. Такие фундаментальные части скелетов схожих объектов близки в метрическом смысле, но нельзя гарантировать схожести их структур: при деформациях протяженных фрагментов границы, которые определяют фундаментальную часть скелета, меняется их положение друг относительно друга, и, значит, может меняться структура фундаментальной части. Отметим, что ветви фундаментальной части скелетов можно разделить на два типа: основные

Работа выполнена при поддержке РФФИ, проекты № 11-07-00462 и № 11-01-00783.

и связующие. На основные ветви изменения границы в пределах точности существенно не влияют, а связующие ветви определяют способ соединения основных ветвей и могут появляться и исчезать при деформациях границы.

Существует множество методов выделения фундаментальной части скелета, основанных на так называемой «стрижке» скелета [2]. Идея большинства таких методов состоит в следующем: после удаления ветви по полученному скелету восстанавливается силуэт и сравнивается с исходным объектом. По результату сравнения на основе различных правил делается вывод, является ли удаляемая ветвь шумовой.

Помимо методов «стрижки», существует другой подход, основанный на выделении в скелете такого подмножества, которое будет устойчиво к изменению границы. Например, в работе [3] вводится понятие «инъективной» области — замкнутой ограниченной области, граница которой не имеет выпуклых углов, а все максимальные вписанные круги касаются границы более, чем в двух точках. В работе показано, что если инъективная область аппроксимирует некоторую замкнутую ограниченную область с точностью $\varepsilon < \min\left\{\rho \tan^2 \frac{\theta}{2}, \frac{\rho}{2}\right\}$ (здесь ρ — наименьший радиус максимального вписанного круга инъективной области, а θ — наименьший угол между радиусами максимального вписанного круга, проходящими через точки касания) в смысле расстояния Хаусдорфа, то Хаусдорфово отклонение скелета аппроксимируемой области от скелета инъективной области не превышает $\frac{\rho\varepsilon}{\rho \sin^2 \frac{\theta}{2} - \varepsilon \cos^2 \frac{\theta}{2}}$. Таким образом, скелет инъективной области можно рассматривать как модель фундаментальной части скелета при относительно небольших «шевелениях» границы. Отметим, что авторы не описывают, каким образом нужно строить инъективную область.

В работе [4] рассматривается область в R^n и подмножество ее скелета M_λ — множество точек скелета, для которых радиус соответствующего вписанного круга не меньше λ. Показано, что λ-скелет M_λ устойчив к изменениям границы, если расстояние Хаусдорфа между областями $\varepsilon < min\left(\frac{\lambda}{10}, \frac{\lambda^3}{50D^2}, \frac{\lambda^4}{1200D^4}\right)$ (здесь D — диаметр исходной области): расстояние между λ-скелетом исходной области и скелетом близкой области не превосходит $\frac{72D^2}{\lambda^2}\varepsilon$. Модель λ-скелета области авторы строят, используя так называемый скелет Вороного — специальное подмножество ребер диаграммы Вороного некоторого набора точек границы области.

Заметим, что в существующих работах, посвященных созданию устойчивых скелетных моделей, вопрос структурного сходства скелетов практически не затрагивается. В этом отношении особый интерес представляет статья [5], в которой предложен метод построения обобщенного скелетного графа. В терминологии авторов при незначительных изменениях границы может произойти «перехлест» ребер скелета, т. е. близко расположенные вершины «меняются местами». В работе описана процедура «склейки» узлов скелета, идея которой заключается в следующем: каждая ветвь скелета с нетерминальными вершинами, такая, что расстояние между вершинами достаточно мало, считается связующей и заменяется на один узел. При этом узел выбирают так, что восстановленный по полученному скелету силуэт отличается от исходной фигуры не более чем на заданную величину.

В [6, 7] было предложено использовать в качестве устойчивого скелетного представления базовый скелет — подмножество скелета многоугольной фигуры. С базовым скелетом связана особая замкнутая область — так называемое скелетное ядро [8]. Эта область образована совокупностью окрестностей ребер базового скелета (границами этих окрестностей являются отрезки прямых и фрагменты парабол и гипербол, в зависимости от типа ребра), внутрь которых попадают ветви скелетов любых замкнутых односвязных областей,

близких многоугольной фигуре в смысле расстояния Хаусдорфа (рис. 1). Скелетное ядро, в частности, позволяет выделять в базовом скелете основные и связующие ветви скелета на основе свойств их окрестностей.

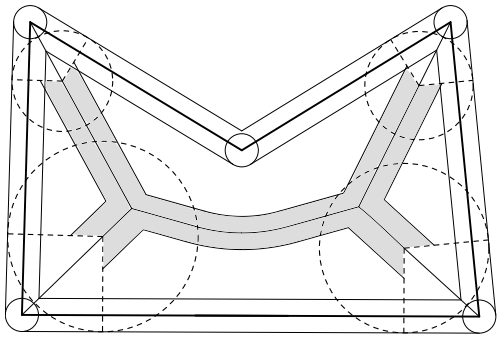


Рис. 1: Базовый скелет и скелетное ядро многоугольной фигуры

С ростом величины точности аппроксимации базовый скелет изменяется монотонно и непрерывно (в смысле расстояния Хаусдорфа). В настоящей работе проводится обоснование этих свойств базового скелета и исследуется процесс изменения. Для описания поведения семейства базовых скелетов, соответствующих различным значениям точности, вводится понятие разметки скелета — множества точек скелета многоугольной фигуры, характеризующего существенные изменения базового скелета и позволяющего строить базовые скелеты для заданного набора или интервала значений точности аппроксимации.

Базовый скелет многоугольной фигуры

Введем необходимые определения в соответствии с [7, 9].

Определение 1. *Фигурой называется связная замкнутая область на плоскости, ограниченная конечным числом непересекающихся жордановых кривых.*

Рассмотрим фигуру F на плоскости R^2 с евклидовым расстоянием $d(p,q), p,q \in R^2$.

Определение 2. *Пустым кругом фигуры F с центром в точке p и радиусом $r \geqslant 0$ называется замкнутое множество точек $\tilde{C}_r(p) = \{q : q \in R^2, d(p,q) \leqslant r\}$ такое, что $\tilde{C}_r(p) \subset F$.*

Определение 3. *Максимальным пустым кругом называется пустой круг, который не содержится ни в каком другом пустом круге.*

Определение 4. *Скелетом фигуры $S(F)$ называется множество центров всех ее максимальных пустых кругов.*

Скелет многоугольной фигуры представляет собой конечное множество отрезков прямых и фрагментов парабол: для двух вершин вогнутых углов u_1 и u_2 границы серединной осью является серединный перпендикуляр к отрезку u_1u_2; для двух сегментов границы — биссектриса угла, образованного этими сегментами; для вогнутой вершины u_1 и сегмента — фрагмент параболы с фокусом u_1 и директрисой, содержащей сегмент. Таким образом, скелет многоугольной фигуры имеет вид плоского графа (рис. 2). Ребрами этого графа являются отрезки и фрагменты парабол, а вершинами следующие точки скелета:

- точки, являющиеся вершинами выпуклых углов границы (эти точки представляют собой терминальные вершины скелета степени 1);

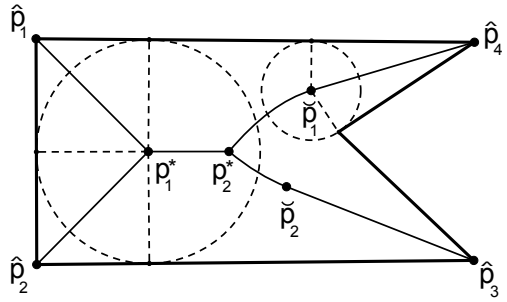


Рис. 2: Вершины и ребра скелета многоугольной фигуры

- точки, максимальные пустые круги с центрами в которых касаются границы в трех и более точках (вершины скелета степени 3 и более);

- точки, которые являются общими точками двух ветвей и имеют максимальный пустой круг, касающийся границы в двух точках (вершины скелета степени 2).

Пусть P — односвязная многоугольная фигура, ε — некоторое неотрицательное число. В качестве расстояния между множествами будем использовать расстояние Хаусдорфа H.

Определение 5. *Круг C называется ε-допустимым кругом для P, если:*
1) $H(P, P \bigcup C) \leqslant \varepsilon$;
2) $H(\partial P, \partial(P \bigcup C)) \leqslant \varepsilon$.

Определение 6. *Круг C называется максимальным ε-допустимым кругом для P, если он является ε-допустимым кругом для P и не содержится целиком ни в каком другом ε-допустимом для P круге.*

Справедливы следующие утверждения.

Утверждение 1. *Если $C_r(p)$ – максимальный ε-допустимый круг для P, то $r \geqslant \varepsilon$.*

Утверждение 2. *Если $C_r(p)$ – максимальный ε-допустимый круг для P, то $C_{r-\varepsilon}(p)$ – максимальный пустой круг для P.*

Утверждение 3. *Если $C_r(p)$ — максимальный пустой круг для P, то $C_{r+\varepsilon}(p)$ — максимальный ε-допустимый круг для P.*

Следствием этих утверждений является следующая

Теорема 1. *Множество центров максимальных ε-допустимых кругов для P совпадает со множеством центров максимальных пустых кругов для P.*

Пусть C — максимальный ε-допустимый круг для P. Точки, в которых соответствующий C максимальный пустой круг C' касается границы фигуры, разбивают границу на фрагменты $P_1, P_2, \ldots, P_n$, $n \geqslant 2$, а радиусы круга C, проходящие через эти точки, разбивают окружность круга C на дуги $L_1, L_2, \ldots, L_n$ (рис. 3).

Определение 7. *Максимальный ε-допустимый круг C называется базовым кругом для многоугольной фигуры P, если $\exists i, j : i \neq j, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n$ такие, что $H(P_i, L_i) \geqslant \varepsilon$ и $H(P_j, L_j) \geqslant \varepsilon$.*

Определение 8. *Базовым скелетом $S_{base}(P, \varepsilon)$ многоугольной фигуры P называется множество центров всех базовых кругов области.*

Из теоремы 1 следует, что базовый скелет P является подмножеством скелета P.

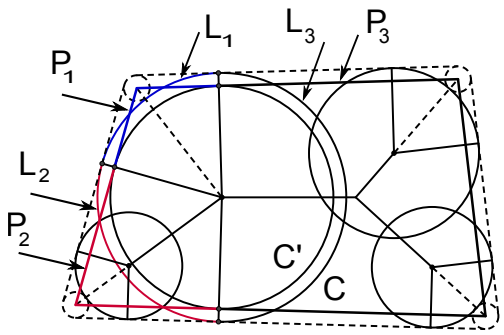


Рис. 3: Фрагменты границы фигуры и дуги окружности максимального ε-допустимого круга

Монотонность изменения базового скелета

Рассмотрим, как изменяется базовый скелет при росте величины точности аппроксимации ε. Точки скелета могут быть трех типов: терминальные вершины (они совпадают с вершинами границы), нетерминальные вершины и внутренние точки ребер. Для любой точки скелета определен максимальный пустой круг с центром в этой точке. Для терминальных вершин точка касания соответствующего максимального пустого круга единственна и совпадает с самой вершиной; для внутренних точек ребер максимальный пустой круг касается границы в двух точках; для нетерминальных вершин скелета — в $k \geqslant 2$ точках. Таким образом, для каждой точки O скелета граница фигуры разбивается точками касания максимального пустого круга с центром в O на $n \geqslant 2$ фрагментов (в случае терминальной вершины единственную точку касания тоже считаем фрагментом границы). Для фиксированного ε рассмотрим максимальный ε-допустимый круг с центром в O. Радиусы, проведенные через точки касания максимального пустого круга, разбивают окружность максимального ε-допустимого круга на n дуг, соответствующих фрагментам границы. Для каждой пары соответствующих множеств «дуга-фрагмент границы» определено расстояние Хаусдорфа между ними. Если существует две пары таких множеств, для которых расстояние Хаусдорфа между элементами пары больше либо равно ε, то точка O принадлежит базовому скелету. Поскольку граница представляет собой замкнутую ломаную, то для вычисления расстояния Хаусдорфа между элементами пары «дуга-фрагмент границы» достаточно знать расстояния от дуги до вершин фрагмента границы.

Исследуем, как изменяется базовый скелет в зависимости от точности аппроксимации ε. При ε=0 все точки скелета являются базовыми. При увеличении ε процесс «выпадения» точек из базового скелета начнется от терминальных вершин скелета. Пусть p – некоторая точка скелета, C' – максимальный пустой круг с центром в точке p радиуса r, C – максимальный ε-допустимый круг с центром в точке p и радиусом $r + \varepsilon$.

Обозначим U_i, $i = 1, \ldots, n$ — подмножества вершин границы, принадлежащих фрагментам, на которые разбивается граница точками касания круга C'. Рассмотрим максимальное расстояние от точки p до точек из множества U_i:

$$d_i = max\{d(p, u))|u \in U_i\}.$$

Упорядочим расстояния d_i, $i = 1, \ldots, n$, по возрастанию:

$$d_1 \leqslant d_2 \leqslant \ldots \leqslant d_{n-1} \leqslant d_n,$$

и выберем такое подмножество U_j, что соответствующее расстояние d_j является вторым по величине. Если таких подмножеств несколько, рассмотрим любое из них. Если же $d_1 =$

$= d_2 = \cdots = d_{n-1} = d_n$, то выберем любое из подмножеств U_i. В дальнейшем выбранное подмножество вершин границы будем обозначать U', а соответствующее максимальное расстояние от точки p до точек U' будем обозначать d'.

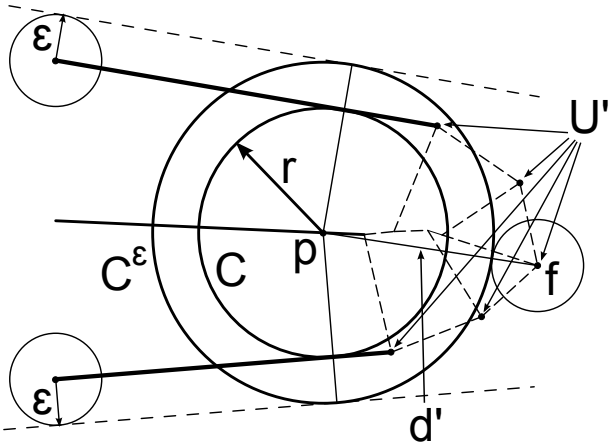


Рис. 4: Терминальная вершина базового скелета

Теорема 2. *Базовый скелет односвязной многоугольной фигуры монотонно зависит от точности аппроксимации ε.*

Доказательство. При $\varepsilon > (d' - r)/2$ (рис.4) точка p не будет базовой, так как нарушается условие Определения 7. Значит, существует такое ε, при котором точка p «выпадает» из базового скелета. Пусть при $\varepsilon = \varepsilon_1$ максимальный ε-допустимый круг C^{ε_1} с центром в точке p не базовый. Докажем, что для любого $\varepsilon_2 > \varepsilon_1$ соответствующий максимальный ε_2-допустимый круг C^{ε_2} с центром в p также не является базовым. Так как круг C^{ε_1} не базовый, то для $k \geqslant n-1$ дуг окружности C^{ε_1} расстояние Хаусдорфа между дугой и соответствующим фрагментом границы меньше ε_1. Рассмотрим любую из таких дуг. Обозначим эту дугу L_1, а соответствующий фрагмент границы $P - P_{12}$. Соответственно, $H(L_1, P_{12}) < \varepsilon_1$. Пусть L_2 - дуга окружности C^{ε_2}, образуемая радиусами C^{ε_2}, проходящими через те же точки касания соответвующего максимального пустого круга, что и радиусы C^{ε_1}, образующие дугу L_1. Поскольку для расстояния Хаусдорфа выполняется неравенство треугольника $H(L_2, P_{12}) \leqslant H(L_2, L_1) + H(L_1, P_{12})$, то $H(L_2, P_{12}) < (\varepsilon_2 - \varepsilon_1) + \varepsilon_1 = \varepsilon_2$. Получаем, что для $k \geqslant n-1$ дуг окружности C^{ε_2} расстояние Хаусдорфа между дугой и соответствующим фрагментом границы меньше ε_2, т. е. круг C^{ε_2} тоже не является базовым. Это означает, что если ε достигло значения, при котором точка перестает быть базовой, то при всех последующих значениях ε эта точка также не будет принадлежать базовому скелету.

Из теоремы следует, что базовый скелет, соответствующий точности $\varepsilon_2 > \varepsilon_1$, является подмножеством базового скелета точности ε_1.

«Стирание» скелета

Итак, при $\varepsilon = \dfrac{d' - r}{2}$ точка p является терминальной вершиной базового скелета. Посмотрим, как происходит изменение базового скелета при росте ε.

Как известно, ребро скелета многоугольной фигуры может быть трех типов: отрезок, порожденный парой сегментов границы; отрезок, порожденный парой вершин границы;

фрагмент параболы, порожденный вершиной и сегментом. Рассмотрим, как ведет себя терминальная точка базового скелета в каждом из этих случаев.

Пусть s_1 и s_2 — два сегмента границы, точка p принадлежит бисектору этой пары, $\varepsilon > 0$ такое, что точка p является терминальной точкой базового скелета (рис.5). Пусть

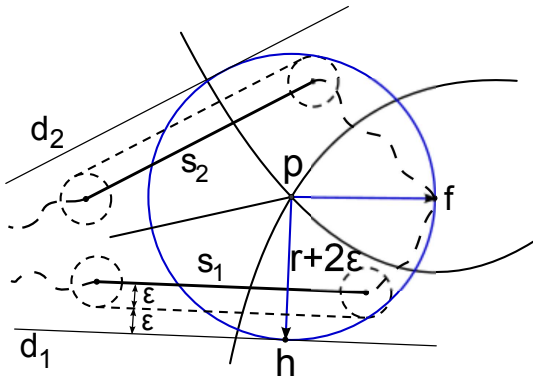


Рис. 5: Стирающие кривые для ребра — бисектора двух сегментов границы

f — наиболее удаленная от точки p вершина границы из множества U', $r_\varepsilon = r + \varepsilon$ — радиус базового круга с центром в точке p. Рассмотрим окружность радиуса $r + 2\varepsilon$ с центром в p. Нетрудно видеть, что эта окружность проходит через точку f. Пусть ph — радиус этой окружности, перпендикулярный сегменту s_1. Так как $pf = ph$, то точка p лежит на параболе с фокусом f и директрисой, проходящей через точку h и параллельной сегменту границы s_1 (аналогично рассуждая, видим, что точка p принадлежит параболе с фокусом f и директрисой, параллельной сегменту s_2 и лежащей на расстоянии 2ε от s_2). Таким образом, терминальная точка базового скелета лежит на пересечении параболы и ребра скелета. Увеличим ε на некоторую достаточно малую величину δ. Пусть p_1 — терминальная вершина базового скелета точности $\varepsilon_1 = \varepsilon + \delta$, r_{ε_1} — радиус базового круга с центром в p_1. Тогда $r_{\varepsilon_1} = r_\varepsilon + \delta$. Точка p_1 будет точкой пересечения ребра скелета и параболы с фокусом в той же точке f. Директриса этой параболы параллельна сегменту s_1 и лежит на расстоянии $2(\varepsilon + \delta)$ от него. В системе координат с центром в фокусе f и осью абсцисс, параллельной сегменту s_1, параболы при разной точности ε имеют один вид:

$$y = \frac{x^2}{4(\varepsilon + c)} - (\varepsilon + c)$$

Отсюда видим, что при увеличении ε директриса удаляется от фокуса, ветви параболы «расходятся» и ребро скелета «стирается» точкой пересечения с параболой.

Рассмотрим теперь случай, когда элементами, порождающими ребро, являются сегмент s и вершина a границы (рис. 6). Рассуждая аналогично, получим для сегмента s параболу с фокусом f и директрисой, параллельной s и лежащей на расстоянии 2ε от s. Рассмотрим базовый круг с центром в p и радиусом r_ε. Очевидно, что $r_\varepsilon = pf - \varepsilon = pa + \varepsilon$. Следовательно, $pf - pa = 2\varepsilon$, т. е. разность расстояний постоянна. Значит, точка p лежит на гиперболе с фокусами f и a и расстоянием между вершинами 2ε.

Теперь рассмотрим ситуацию, когда ребро является бисектором двух вершин границы a и b (рис. 7).

Нетрудно видеть, что в данном случае центр базового круга лежит на пересечении ветвей двух гипербол — с фокусами $\{a, f\}$ и $\{b, f\}$ соответственно. Расстояние между вершинами у гипербол одинаково и равно 2ε.

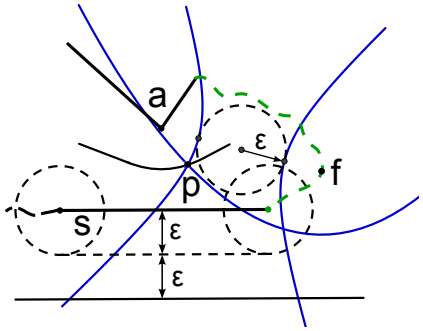


Рис. 6: Стирающие кривые для ребра — бисектора сегмента и вершины границы

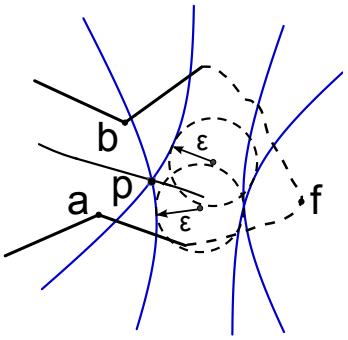


Рис. 7: Стирающие кривые для ребра — бисектора двух вершин границы

Мы рассмотрели стирание ребра скелета в трех элементарных случаях. Однако возможно такое взаимное расположение вершин границы, когда процесс изменения скелета более сложен. Исследуем такие ситуации подробнее.

Точки смены стирающих кривых ребра

Итак, стирающую кривую для ребра определяют инцидентные элементы границы и самая удаленная точка f из подмножества вершин границы U'. Для определения положения точки f воспользуемся диаграммой Вороного дальней точки [10]. Для каждой вершины границы из подмножества U' определим «зону дальности» — множество точек, расстояние до которых от этой вершины больше, чем от любой другой. Таким образом, в «зону дальности» точки f попадут ребра скелета, для которых эта точка является наиболее удаленной из подмножества вершин U'. Если ребро скелета целиком лежит в одной зоне дальности, то для всех точек ребра вершина, определяющая стирающие кривые, единственна. В противном случае ребро разбивается на несколько фрагментов, каждому из которых соответствует своя вершина (рис. 8). Отметим, что диаграмму дальней точки имеет смысл строить для подмножества, состоящего только из выпуклых вершин границы, так как невыпуклая вершина не может быть самой удаленной точкой для ребра.

Для ребра, которое пересекается с ребром диаграммы Вороного дальней точки, стирание происходит следующим образом (рис. 9). Пусть ребро v_1v_2 пересекает ребро диаграммы в точке q. Это значит, что для точек фрагмента ребра v_1q самой удаленной точкой является вершина a, поэтому v_1q стирается парой кривых, фокусом (или одним из фокусов) которых является точка a (в рассматриваемом примере это парабола с фокусом a и гипербола, один из фокусов которой a). В точке q происходит смена самой удаленной точки и, следовательно, пары кривых. Соответственно, фрагмент ребра qv_2 стирается параболой

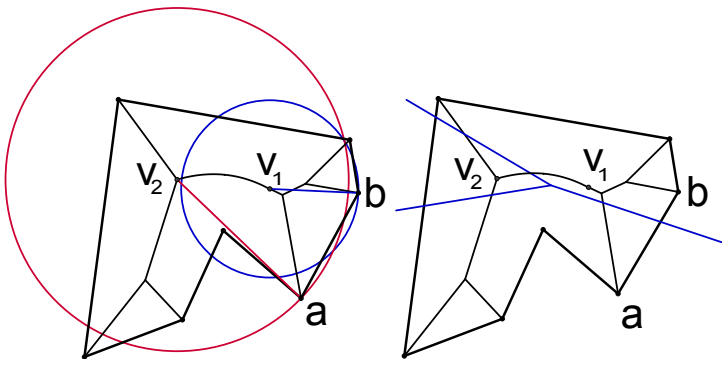


Рис. 8: Фрагменты ребра, для которых наиболее удаленные точки из U' различны

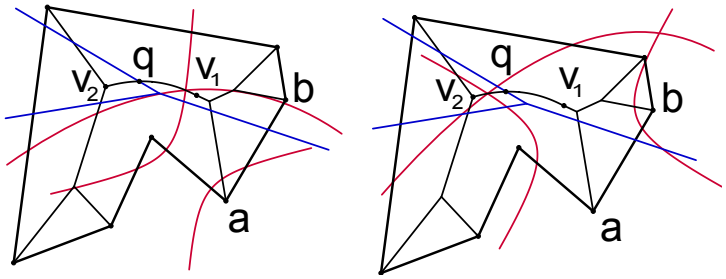


Рис. 9: Точка смены стирающих кривых ребра

с фокусом b и гиперболой, один из фокусов которой b. Заметим, что ребро стирается в одном направлении.

Возможна и более сложная ситуация. Рассмотрим фрагмент фигуры на рис.10. Будем считать, что рассматриваемое подмножество вершин границы является вторым по дальности подмножеством U'.

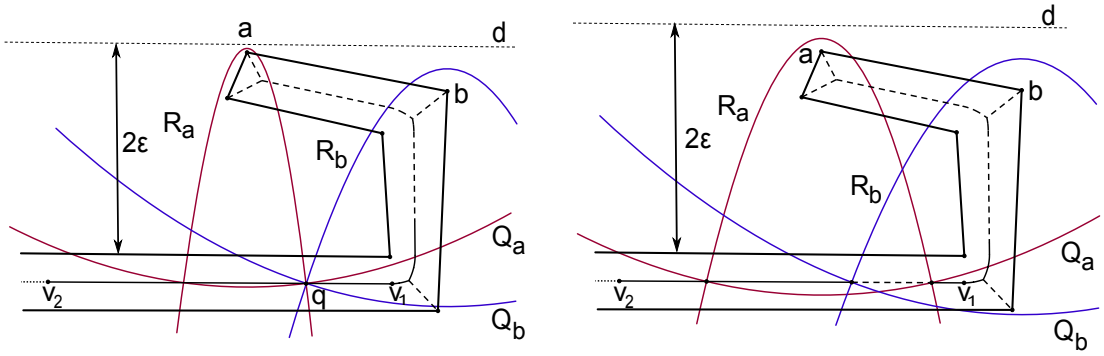


Рис. 10: Нарушение связности базового скелета в точке смены стирающих кривых

Рассмотрим ребро v_1v_2. На нем находится точка q, в которой ребро пересекается с ребром диаграммы Вороного дальней точки, и, соответственно, происходит смена стирающей кривой. Эта точка равноудалена от вершин a и b границы. Стирающими кривыми для ребра v_1v_2 являются две пары парабол: (R_a, Q_a) для части ребра v_1q и (R_b, Q_b) для части ребра qv_2. Рассмотрим параболы с одинаковой директрисой d R_a и R_b. При достаточно малых значениях ε директриса d лежит между фокусом параболы и сегментом границы, которому параллельна директриса. При значениях ε, больших некоторого ε_1, фокус и сегмент границы лежат по одну сторону от директрисы. Это значит, что параболы пе-

ресекают ребро v_1v_2. Таким образом, с точки q начнется стирание ребра v_1v_2 параболами R_a и R_b в разных направлениях — т. е. в этой точке произойдет нарушение связности и базовый скелет разделится на две части.

Отметим, что значение ε, при котором стирающие кривые проходят через точку q, меньше, чем соответствующие значения для концевых точек ребра (так как точка q «выпадет» из базового скелета раньше, чем точки v_1 и v_2).

Точки касания ребра и стирающей кривой

Исследуем ситуации, в которых одна и та же пара стирающих кривых пересекает ребро в нескольких точках и стирает его в разных направлениях.

Рассмотрим сначала случай, когда ребро скелета представляет собой фрагмент параболы, а стирают ребро, соответственно, парабола и гипербола. Директрисы парабол параллельны, но совпадать параболы не могут, так как их фокусы всегда различны: фокус параболы, определяющей ребро, не является выпуклой вершиной, поэтому он не может быть наиболее удаленной точкой для ребра, и, следовательно, фокусом стирающей параболы. При $\varepsilon = 0$ директрисы парабол совпадают. При увеличении ε директриса стирающей параболы удаляется от директрисы параболы, определяющей ребро. Поэтому возможна лишь ситуация, когда с ростом ε параболы сначала пересекаются в двух точках, а затем касаются друг друга (при дальнейшем увеличении ε параболы не имеют точек пересечения) (рис. 11). В момент, когда параболы касаются друг друга, стирающая гипербола вырождается в лучи, исходящие из фокусов. Это происходит при ε, равном половине расстояния между фокусами.

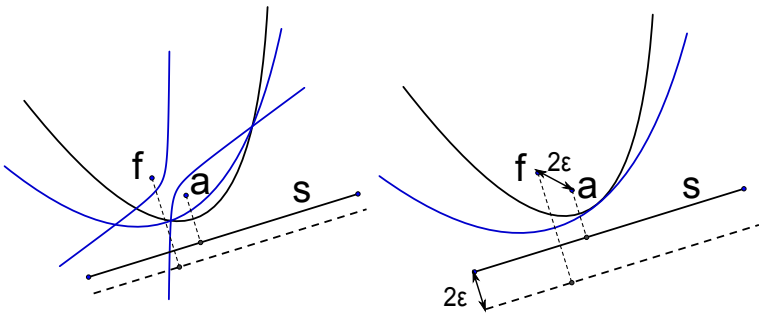


Рис. 11: Касание параболического ребра и стирающей параболы

Таким образом, параболическое ребро будет стираться с двух сторон, и последней точкой ребра будет точка касания парабол. На рис. 12 показан фрагмент фигуры, для которого стирание скелета закончится описанным образом.

Теперь рассмотрим случай, когда ребром скелета является отрезок, порожденный двумя вершинами границы (рис. 13). Здесь рассуждения аналогичны.

Пусть a и b — вершины границы, v_1v_2 — ребро скелета, порожденное вершинами a и b (отрезок), f — самая удаленная точка. Аналогично предыдущему случаю, ребро v_1v_2 стирается парой гипербол с двух сторон. Целиком ребро сотрется, когда одна из гипербол будет касаться отрезка-ребра, а вторая выродится в пару лучей, исходящих из фокусов.

При этом значение ε будет равно половине расстояния между фокусами гиперболы, которая вырождается в лучи. На рис. 14 показан пример фрагмента базового скелета фигуры, оставшегося в результате нарушения связности. Стирание скелета этого фрагмента заканчивается на ребре — бисекторе пары вершин границы.

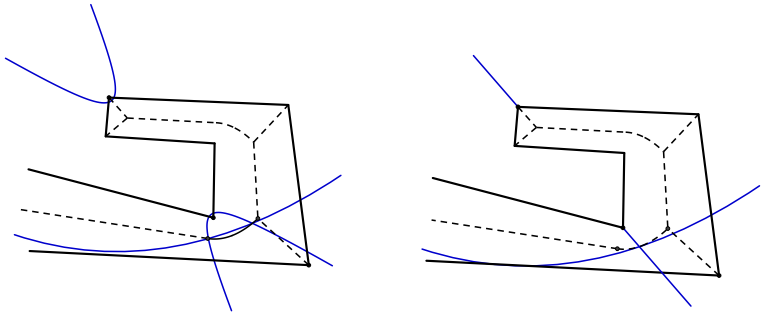


Рис. 12: Пример скелета фигуры с точкой касания на параболическом ребре

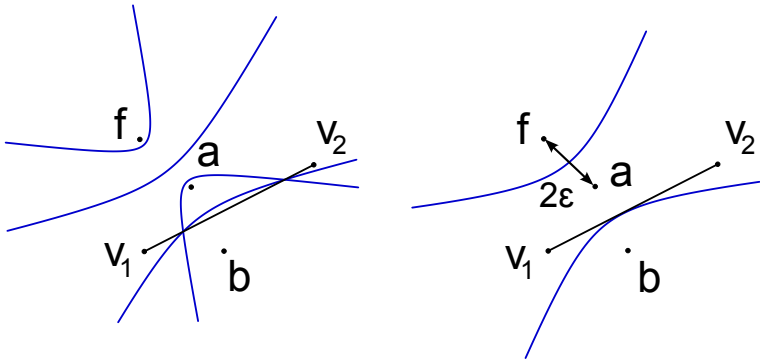


Рис. 13: Касание ребра — бисектора двух вершин и стирающей гиперболы

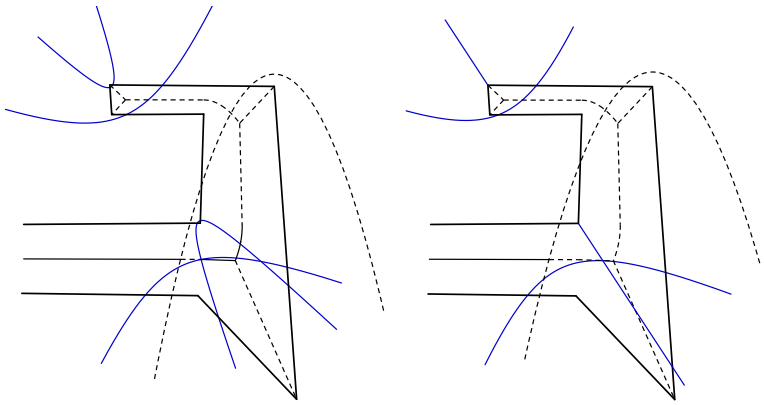


Рис. 14: Пример скелета фигуры с точкой касания на ребре — бисекторе двух вершин

Возможна ли аналогичная ситуация в случае, если оба элемента являются сегментами? Директриса секущей параболы разбивает плоскость на две полуплоскости. Если фокус параболы (наиболее удаленная точка) f и ребро скелета лежат в разных полуплоскостях, то парабола не пересекает ребро. Пусть точка f и ребро скелета лежат в одной полуплоскости. Тогда при увеличении ε расстояние от ребра скелета до директрисы параболы будет увеличиваться. Следовательно, ситуация, когда парабола пересекает ребро в двух точках при некотором ε_1, а при некотором $\varepsilon_2 > \varepsilon_1$ касается ребра, невозможна. Но возможно обратное: при увеличении ε одна из стирающих парабол сначала касается ребра, а затем пересекает его в двух точках. При этом вторая стирающая парабола в момент касания вырождена и представляет собой луч из точки f, перпендикулярный соответствующему сегменту границы и пересекающий ребро в точке касания ребра первой параболой. Касание происходит при ε, равном- половине расстояния от точки f до сегмента границы,

определяющего вырожденную в луч параболу (т. е. в момент, когда директриса, параллельная сегменту, проходит через точку f) (рис. 15). Соответственно, в этом случае, в отличие от двух предыдущих, нарушается связность базового скелета.

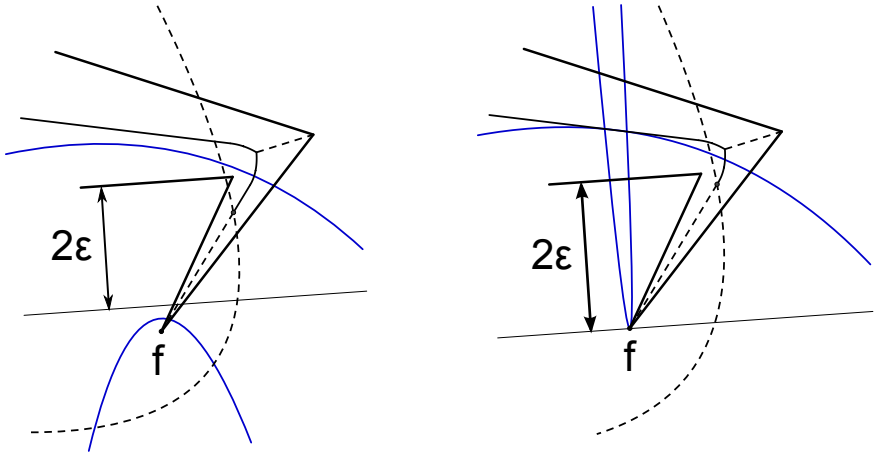


Рис. 15: Пример скелета фигуры с точкой касания на ребре – бисекторе двух сегментов

Центральные точки скелета

При достижении некоторого значения ε из базового скелета исчезнут все ребра. Возникает вопрос, какая точка скелета исчезнет последней. Логично ожидать, что в последней точке v_0 «сойдутся» две пары стирающих кривых, порожденные разными множествами U'. Пусть v_0 является внутренней точкой некоторого ребра v_1v_2, для отрезков v_1v_0 и v_0v_2 множества U' различны и каждый из этих отрезков стирается в одном направлении. Тогда точка v_0 на ребре v_1v_2 равноудалена от дальних точек $f_1 \in U'_1$ и $f_2 \in U'_2$. Через эту точку проходят две пары кривых, одна из которых соответствует множеству U'_1, а другая — множеству U'_2 и в ней закончится процесс стирания скелета (рис. 16). Это предположение можно обобщить на случай, когда точка v_0 является вершиной скелета: через вершину проходит n пар различных стирающих кривых, где n — степень вершины.

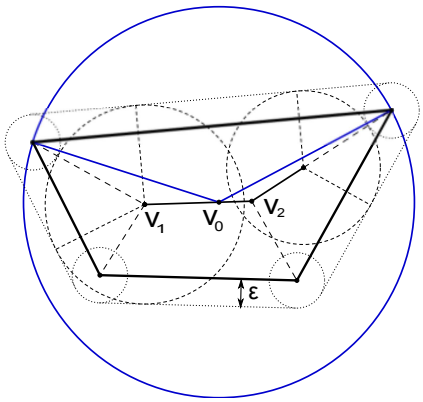


Рис. 16: Центральная точка скелета

Определение 9. *Точка скелета фигуры называется центральной, если максимальный пустой круг с центром в ней разбивает границу точками касания на n фрагментов и эта точка равноудалена от каждой из дальних точек $f_i \in U'_i, i = 1, ..., n$.*

Возможна и ситуация, когда центральная точка является точкой нарушения связности. На рис. 17 приведен пример симметричной фигуры с такой точкой: в этой точке происходит разделение скелета на две части, стирание каждой из которых заканчивается в точке касания параболического ребра и стирающей кривой. Отметим, что в данном случае все точки ребра, на котором лежит точка нарушения связности, являются центральными.

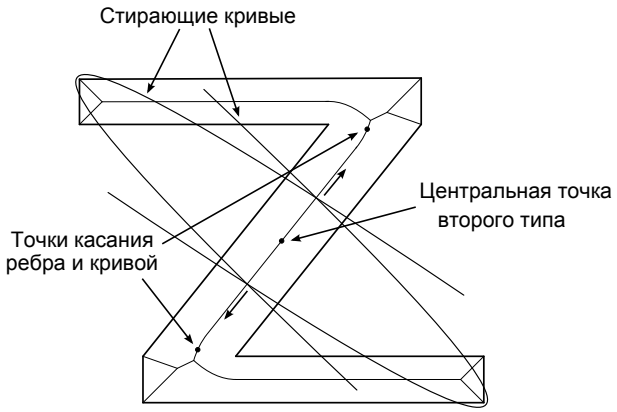


Рис. 17: Центральная точка скелета 2-го типа

Определение 10. *Центральная точка скелета фигуры называется центральной точкой 1-го типа, если в ней заканчивается стирание ребра. Центральная точка, в которой происходит нарушение связности, называется центральной точкой 2-го типа.*

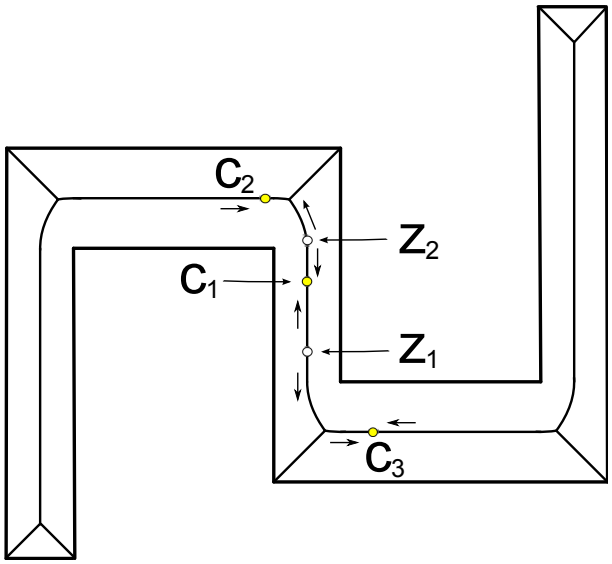


Рис. 18: Пример скелета фигуры с тремя центральными точками 1-го типа

Заметим, что в случае наличия в скелете точек нарушения связности центральных точек 1-го типа может быть несколько. На рис. 18 изображена фигура, при стирании скелета которой произойдет нарушение связности в двух точках z_1 и z_2, в результате чего базовый скелет разделится на три фрагмента. Каждый из этих фрагментов содержит свою центральную точку 1-го типа (c_1, c_2, c_3), в которой закончится стирание фрагмента.

Разметка скелета

Итак, при росте ε ребра базового скелета стираются парами кривых — парабол и гипербол. Состав пары и положение стирающих кривых для каждого ребра v зависит от типа и положения порождающих ребро элементов границы, наиболее удаленной от v вершины $f \in U'$ и величины ε. При этом ребро скелета может стираться несколькими парами кривых, поскольку вершина f может быть различной для разных фрагментов ребра v. На ребре может находиться точка касания ребра и стирающей кривой, в которой заканчивается либо начинается (в зависимости от типа порождающих ребро элементов границы) стирание ребра скелета одной парой кривых в противоположных направлениях. Отдельный тип представляют собой центральные точки, в которых заканчивается или начинается стирание ребра двумя парами кривых, порожденными разными множествами вершин границы.

Определение 11. *Разметка скелета — это множество точек скелета, в которое входят:*

– вершины скелета;

– точки смены пары стирающих кривых;

– точки касания стирающей кривой и ребра;

– центральные точки скелета 1-го и 2-го типов.

При этом каждой точке множества сопоставлен набор значений точности $\{\varepsilon_i\}, 1 \leqslant i \leqslant n$ *(для вершины скелета* n *равно степени этой вершины, для внутренней точки ребра* n *равно 2), при которых соответствующие стирающие кривые проходят через данную точку.*

Будем рассматривать точки разметки как вершины скелета. В результате получим размеченный скелет (рис. 19), каждое ребро которого стирается одной парой кривых в одном направлении. При этом с каждым ребром связаны два значения точности ε, при которых стирающая пара проходит через концевые точки ребра.

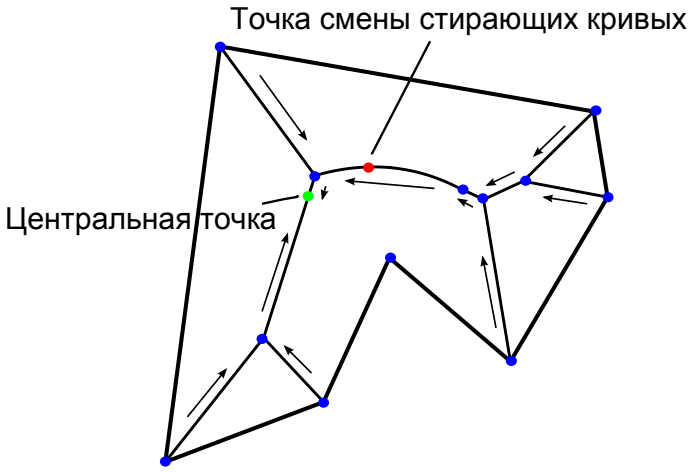


Рис. 19: Размеченный скелет

Непрерывность изменения базового скелета

Исследуем процесс движения точки пересечения стирающей кривой и ребра скелета. Сначала рассмотрим случай, когда ребром скелета является отрезок, порожденный парой сегментов границы, а стирающими кривыми — две параболы. Будем рассматривать систему координат с центром в фокусе стирающей параболы и осью абсцисс, параллельной директрисе параболы и лежащей на расстоянии $r + 2\varepsilon$ от нее, где r — расстояние от

сегмента границы до наиболее удаленной точки (в рассматриваемой системе — начала координат). Тогда уравнение стирающей параболы можно записать в виде:

$$y + \frac{r + 2\varepsilon}{2} = \frac{x^2}{2(r + 2\varepsilon)}$$

Обозначим отрезок, являющийся ребром скелета, $[(x_1, y_1), (x_2, y_2)]$. Уравнение прямой, проходящей через концевые точки ребра, запишем в виде $y = kx + b$, где $k = \dfrac{y_2 - y_1}{x_2 - x_1}$, $b = y_1 - kx_1$. (Заметим, что в выбранной системе координат $x_1 \neq x_2$, так как отрезок $[(x_1, y_1), (x_2, y_2)]$ не может быть перпендикулярен порождающему его сегменту границы, а значит, и директрисе параболы.) Координата x точки пересечения параболы и отрезка — это решение уравнения

$$x^2 - 2tkx - t^2 - 2tb = 0 \tag{1}$$

где $t = r + 2\varepsilon$. Корни этого уравнения имеют вид:

$$x^{\pm}(t) = tk \pm \sqrt{t^2(k^2 + 1) + 2tb} \tag{2}$$

Парабола пересекает отрезок при $t \geqslant -\frac{2b}{k^2+1}$ Очевидно, что $x^{\pm}(t)$ непрерывны. Координата y точки пересечения линейно зависит от x, и, следовательно, точки пересечения движутся по прямой непрерывно.

Рассмотрим случай, когда ребром скелета является отрезок, порожденный парой вершин границы, а парой стирающих кривых являются две гиперболы. Будем рассматривать систему координат, такую, что прямая, проходящая через фокусы гиперболы совпадает с осью Ox, а центр системы координат лежит в середине отрезка, соединяющего фокусы. Тогда уравнение гиперболы имеет вид:

$$\frac{x^2}{\varepsilon^2} - \frac{y^2}{r^2 - \varepsilon^2} = 1 \tag{3}$$

где r — половина расстояния между фокусами, $0 < \varepsilon < r$. Уравнения асимптот этой гиперболы имеют вид:

$$y = \pm\frac{\sqrt{r^2 - \varepsilon^2}}{\varepsilon}x$$

Уравнение прямой, проходящей через точки (x_1, y_1) и (x_2, y_2), запишем в виде $y = kx + b$ (ситуацию когда $x_1 = x_2$ рассмотрим отдельно). Тогда уравнение для координаты x точки пересечения гиперболы с прямой имеет вид:

$$(r^2 - \varepsilon^2 - \varepsilon^2k^2)x^2 - 2kb\varepsilon^2x - b^2\varepsilon^2 - \varepsilon^2(r^2 - \varepsilon^2) = 0 \tag{4}$$

Если $r^2 - \varepsilon^2 - \varepsilon^2k^2 \neq 0$, то корни уравнения можно записать в виде:

$$x^{\pm} = \frac{kb\varepsilon^2 \pm \varepsilon\sqrt{(r^2 - \varepsilon^2)(b^2 + r^2 - \varepsilon^2 - \varepsilon^2k^2)}}{r^2 - \varepsilon^2 - \varepsilon^2k^2} \tag{5}$$

Нетрудно видеть, что при $\varepsilon > \sqrt{\dfrac{r^2 + b^2}{k^2 + 1}}$ гипербола не пересекает прямую, при $\varepsilon = \sqrt{\dfrac{r^2 + b^2}{k^2 + 1}}$ прямая касается гиперболы, при $\varepsilon < \sqrt{\dfrac{r^2 + b^2}{k^2 + 1}}$ гипербола пересекает прямую в двух точках.

При $r^2 - \varepsilon^2 - \varepsilon^2 k^2 = 0$, т. е. при значении точности

$$\varepsilon^* = \frac{r}{\sqrt{k^2+1}} \tag{6}$$

квадратное уравнение (4) обращается в линейное, решение которого имеет вид:

$$x^* = \frac{-r^2k^2 - b^2 - b^2k^2}{2kb(k^2+1)} \tag{7}$$

Геометрический смысл этого случая заключается в том, что прямая параллельна одной из асимптот гиперболы.

Пусть $k \neq 0, b \neq 0$. Рассмотрим движение точек пересечения для случая $k > 0, b < 0$ (в остальных случаях рассуждения аналогичны) (рис. 20). Обозначим угол наклона прямой $\alpha \in (0, \pi/2)$, угол наклона асимптоты, лежащей в первой четверти, $\beta \in (0, \pi/2)$, а точки пересечения $p_1 = (x_{p_1}, y_{p_1})$ и $p_2 = (x_{p_2}, y_{p_2})$. Если $0 < \varepsilon < \varepsilon^*$, то $\alpha < \beta$, $x_{p_2} = x^- < 0 < x_{p_1} = x^+$. При увеличении ε расстояние между вершинами гиперболы увеличивается, а угол β между асимптотой и осью Ox уменьшается. При этом значение $x_{p_1} = x^+$ увеличивается, а $x_{p_2} = x^-$ уменьшается. Когда значение β достигает значения α (и асимптота становится параллельна прямой), прямая лежит ниже асимптоты и, следовательно, не имеет пересечения с левой ветвью гиперболы. Это значит, что при точности ε^* существует только одна точка пересечения с координатой $x_{p_1} = x^*$.

Нетрудно показать, что при $\varepsilon \to \varepsilon^*$ $x^+ \to x^*$. Рассмотрим x^+ в (5) как отношение двух функций от ε:

$$x^+(\varepsilon) = \frac{f(\varepsilon)}{g(\varepsilon)}$$

Тогда $\lim\limits_{\varepsilon \to \varepsilon^*} g(\varepsilon) = 0$, $\lim\limits_{\varepsilon \to \varepsilon^*} f(\varepsilon) = 0$. Следовательно,

$$\lim_{\varepsilon \to \varepsilon^*} x^+(\varepsilon) = \lim_{\varepsilon \to \varepsilon^*} \frac{f'(\varepsilon)}{g'(\varepsilon)} = \frac{-k^2r^2 - b^2 - k^2b^2}{2kb(k^2+1)} = x^*$$

При дальнейшем уменьшении β снова возникнет вторая точка пересечения p_2, лежащая правее p_1 на той же ветви гиперболы, т. е. $x_{p_2} = x^- > x_{p_1} = x^+ > 0$. При дальнейшем уменьшении β x_{p_1} будет увеличиваться, а x_{p_2} уменьшаться до тех пор, пока их значения не совпадут (в этот момент прямая будет касаться гиперболы).

Таким образом, если заданный отрезок пересекается с левой ветвью гиперболы (лежит в отрицательной полуплоскости по x), то при $\varepsilon < \varepsilon^*$ точка пересечения движется непрерывно, а при $\varepsilon \geqslant \varepsilon^*$ пересечений нет. Если отрезок лежит в положительной полуплоскости, то при $\varepsilon \leqslant \varepsilon^*$ может быть только одна точка пересечения, и она движется непрерывно, а при $\varepsilon > \varepsilon^*$ точек пересечения может быть две, и обе они движутся непрерывно. Координаты x точек пересечения можно записать в виде:

$$x_{p_1} = \begin{cases} x^+, & \text{если } \varepsilon \neq \varepsilon^* \\ x^*, & \text{если } \varepsilon = \varepsilon^* \end{cases}$$

$$x_{p_2} = x^-, \varepsilon \neq \varepsilon^*$$

Если $k \neq 0, b = 0$, то при точности ε^* одна из асимптот совпадет с прямой, следовательно, при $\varepsilon \geqslant \varepsilon^*$ гипербола не будет пересекать прямую. При $\varepsilon < \varepsilon^*$ гипербола пересекает

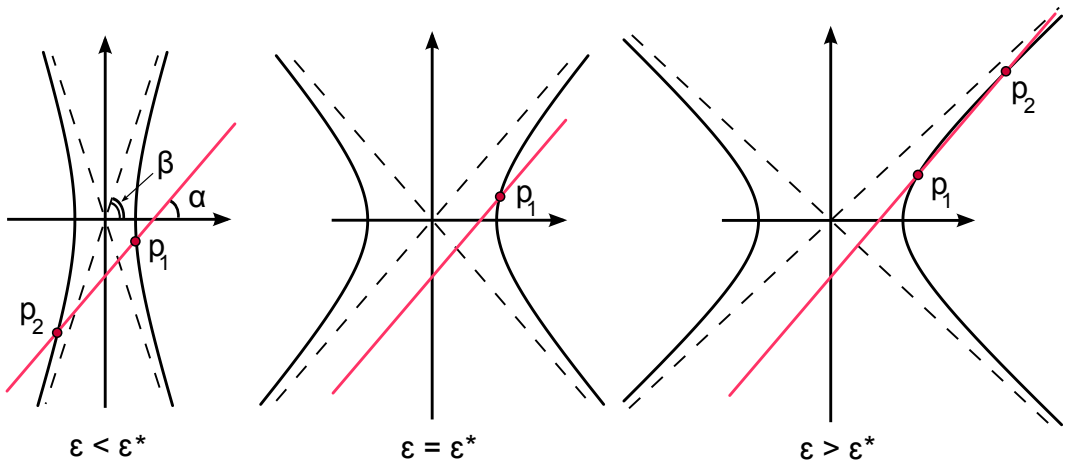


Рис. 20: Точки пересечения стирающей гиперболы и прямой

прямую в двух точках с координатами $x^{\pm}$, лежащих на разных ветвях. Таким образом, и в этом случае точки пересечения движутся непрерывно.

Если $k = 0$, то гипербола также пересекает прямую в двух точках с координатами $x^{\pm}$, лежащих на разных ветвях.

Для случая прямой $x = c$ уравнение координат y точек пересечения имеет вид:

$$y^{\pm} = \pm\frac{\sqrt{(r^2 - \varepsilon^2)(c^2 - \varepsilon^2)}}{\varepsilon}.$$

Здесь при $\varepsilon \leqslant c$ гипербола пересекает прямую, иначе пересечений нет.

Таким образом, движение точки пересечения стирающей гиперболы и ребра носит непрерывный характер.

Теперь рассмотрим движение точки пересечения двух парабол. Напомним, что директрисы парабол параллельны, а фокусы лежат по одну сторону от директрисы скелетной параболы и не совпадают. Будем рассматривать систему координат с центром в фокусе стирающей параболы и осью Ox, параллельной директрисам. Уравнение для координаты x точки пересечения имеет вид:

$$(q - t)x^2 + 2tx_0x - t({x_0}^2 + tq + 2qy_0 - q^2) = 0 \tag{8}$$

где $t = r + 2\varepsilon$, q — расстояние между сегментом и вершиной (x_0, y_0) границы. Отметим, что $q - y_0 = r$. При увеличении значения точности ε директриса стирающей параболы удаляется от фокуса, т. е. увеличивается фокальный параметр t параболы. При этом фокальный параметр q другой параболы остается неизменным. При совпадении значений фокальных параметров $t^* = q$ параболы имеют только одну точку пересечения:

$$x^* = \frac{x_0}{2} + \frac{qy_0}{x_0}. \tag{9}$$

При $t \neq q$ корни уравнения (8) имеют вид:

$$x^{\pm} = \frac{-tx_0 \pm \sqrt{qt(x_0^2 + 2y_0(q - t) - (q - t)^2)}}{(q - t)} \tag{10}$$

Параболы пересекаются при $t \in [r, r+\sqrt{(x_0^2 + y_0^2)}]$. При $t < q$ точки пересечения парабол p_1 и p_2 лежат на разных ветвях скелетной параболы. Предположим, что $x_0 > 0$ (для

случая $x_0 < 0$ рассуждения аналогичны). Тогда $x_{p_2} = x^- < 0 < x_{p_1} = x^+$. При увеличении t x_{p_1} увеличивается, x_{p_2} — уменьшается. Когда значение t достигает значения q, $x_{p_1} = x^*$. Аналогично случаю пересечения гиперболы и прямой, при $t \to q$ $x^+ \to x^*$:

$$\lim_{t \to q} x^+(t) = \lim_{t \to q} \frac{f'(t)}{g'(t)} = \frac{x_0}{2} + \frac{qy_0}{x_0} = x^*.$$

При $t > q$ у парабол снова две точки пересечения, которые сходятся в точку касания парабол при $t \to r + \sqrt{(x_0^2 + y_0^2)}$.

Координаты x точек пересечения парабол p_1 и p_2 можно представить в следующем виде:

$$x_{p_1} = \begin{cases} x^+, & \text{если } t \neq q \\ x^*, & \text{если } t = q \end{cases}$$

$$x_{p_2} = x^-, t \neq q$$

Если $x_0 = 0$, то $t > q$ и $y_0 < 0$. В этом случае параболы пересекаются при $t \leqslant r + |y_0|$ и координаты x точек пересечения изменяются следующим образом:

$$x^{\pm} = \pm \sqrt{\frac{qt(t + 2y_0 - q)}{(q - t)}}.$$

Таким образом, мы показали, что точки пересечения стирающей кривой и ребра скелета движутся непрерывно при изменении ε.

Рассмотрим базовый скелет фигуры P при некотором значении точности ε_1, $S_{base}(P, \varepsilon_1)$. Увеличим значение точности на достаточно малую величину до ε_2. При этом будут стерты фрагменты терминальных ребер базового скелета $S_{base}(P, \varepsilon_1)$. Хаусдорфово расстояние между базовыми скелетами $S_{base}(P, \varepsilon_1)$ и $S_{base}(P, \varepsilon_2)$ будет равно максимальному из расстояний между терминальными вершинами скелетов, лежащих на одном ребре размеченного скелета. Нетрудно видеть, что в силу непрерывного характера движения точек пересечения стирающих кривых с ребрами скелета малое изменение точности ведет к малому изменению расстояния Хаусдорфа между соответствующими базовыми скелетами.

Пусть E — множество значений точности аппроксимации. Из приведенных рассуждений следует

Теорема 3. *Базовый скелет односвязной многоугольной фигуры непрерывно зависит от точности аппроксимации ε в смысле расстояния Хаусдорфа:* $\forall \omega > 0 \quad \exists \sigma > 0 \quad \forall \varepsilon_1, \varepsilon_2 \in E : |\varepsilon_1 - \varepsilon_2| < \sigma \Rightarrow H(S_{base}(P, \varepsilon_1), S_{base}(P, \varepsilon_2)) < \omega$.

Параметрическое семейство базовых скелетов

Таким образом, с многоугольной фигурой связано семейство базовых скелетов, соответствующих различным значениям точности аппроксимации. Размеченный скелет позволяет описать поведение этого семейства в целом и строить базовые скелеты для требуемых значений точности.

Алгоритм построения базового скелета для заданных значений точности аппроксимации состоит в следующем. Для каждого ребра размеченного скелета со значениями точности в концевых точках ε_1 и ε_2 ($\varepsilon_1 < \varepsilon_2$) проверяется выполнение условий:

- если $\varepsilon \leqslant \varepsilon_1$, то ребро целиком принадлежит базовому скелету;
- если $\varepsilon > \varepsilon_2$, то ребро не принадлежит базовому скелету;

- если $\varepsilon_1 < \varepsilon \leqslant \varepsilon_2$, то базовому скелету принадлежит та часть ребра, для точек которой значение точности больше либо равно ε — т. е. от точки пересечения ребра со стирающей кривой при точности ε до концевой точки с точностью ε_2.

С семейством базовых скелетов связано семейство гранично-скелетных моделей [7]. Скелетной частью таких моделей является базовый скелет, а граничная часть представляет собой границу объединения множества всех базовых кругов и отражает те свойства границы, которые являются существенными в пределах точности аппроксимации. Такое семейство является аналогом концепции масштабируемой кривизны границы (curvature scale space) [11] — подхода, основанного на аппроксимации границы кусочно-гладкой кривой, сглаживании этой кривой с помощью фильтра Гаусса и выявлении экстремумов или нулей кривизны границы при разных степенях сглаживания. Для анализа такого семейства используется размеченный скелет, а также оценки значимости для вершин выпуклых углов многоугольника — значения точности аппроксимации, при которых соответствующие вершинам выпуклые особенности перестают быть существенными [12].

Выводы

В работе проведено исследование свойств базового скелета односвязной многоугольной фигуры. Доказана монотонность и непрерывность изменения базового скелета при увеличении величины точности аппроксимации. Для описания процесса изменения базового скелета предложена разметка скелета — множество скелетных точек, соответствующих существенным изменениям базового скелета. Применение устойчивой скелетной модели, основанной на скелете аппроксимирующего многоугольника, позволяет использовать корректные и вычислительно эффективные процедуры скелетизации. Монотонность и непрерывность изменения позволяют рассматривать поведение семейства базовых скелетов в целом и выбирать скелетные модели с нужной точностью аппроксимации. Кроме того, семейство базовых скелетов позволяет строить масштабируемое граничное представление формы, описывающее свойства границы, проявляющиеся при различных степенях детализации, и не требующее применения аппроксимации контура кривыми высших порядков.

Литература

[1] *Blum H.* A transformation for extracting new descriptors of shape // Models for the Perception of Speech and Visual Form, MIT Press, 1967. P. 135–143.

[2] *Shaked D., Bruckstein A.M.* Prunnig medial axes // CVIU. 1998. Vol. 69. No. 2. P. 156–169.

[3] *Choi S., Lee S. W.* Stability Analysis of Medial Axis Transform under Relative Hausdorff Distance. CAVR-TR-99-23, Korea University, 1999.

[4] *Chazal F, Lieutier A.* The λ-medial axis // Graphical Models. 67(4):304–331. July, 2005.

[5] *Domakhina L., Okhlopkov A.* Shape comparison based on skeleton isomorphism // Proceedings of International conference on computer vision theory and applications (VISAPP 2009). Lisbon. Portugal, 2009.

[6] *Местецкий Л. М., Рейер И. А.* Непрерывное скелетное представление изображения с контролируемой точностью // Труды 13 международной конф. ГРАФИКОН-2003. Москва, 2003. С. 246–249.

[7] *Жукова К. В., Рейер И. А.* Параметрическое семейство гранично-скелетных моделей формы // Математические методы распознавания образов: 14-я Всероссийская конференция. Владимирская обл., г. Суздаль, 21–26 сентября 2009 г.: Сборник докладов, с. 346–350.

[8] *Жукова К. В., Рейер И. А.* Структурный анализ формы объекта с помощью скелетного ядра // Интеллектуализация обработки информации: 8-я международная конференция. Республика Кипр, г. Пафос, 17–24 октября 2010 г.: Сборник докладов, с. 350–354.

[9] *Местецкий Л. М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры — Москва: Физматлит, 2009. — 288 с.

[10] *Препарата Ф., Шеймос М.* Вычислительная геометрия: введение — Москва: Мир, 1989. — 478 с.

[11] *Abbasi S., Mokhtarian F., Kittler J.* Curvature scale space image in shape similarity retrieval// MultiMedia Systems. Vol. 7. 1999. P. 467–476.

[12] *Жукова К. В., Рейер И. А.* Параметрический дескриптор формы на основе граничноскелетной модели // Математические методы распознавания образов: 15-я Всероссийская конференция, г. Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов, с. 408–411.

Построение интегрального индикатора в ранговых шкалах с использованием копул для анализа совместного распределения критериев*

Кузнецов М. П.
mikhail.kuznecov@phystech.edu
Московский физико-технический институт

Предложен метод построения интегрального индикатора на основе критериев, выставленных в ранговых шкалах. Для анализа совместного распределения критериев предложено использовать копулы. Предложен алгоритм выбора признаков, основанный на выборе копулы с наибольшим параметром. Работа проиллюстрирована задачей определения статуса редких видов, включенных в Красную книгу РФ.

Ключевые слова: *порядковые шкалы, копула, теорема Скляра, интегральные индикаторы, экспертные оценки.*

Integral indicator construction using copulas*

Kuznetsov M. P.
Moscow Institute of Physics and Technology

We construct an integral indicator of the IUCN Red List of Threatened species. Method of an integral indicator construction based on copulas which describe statistical bounds between the features. We propose a two-step algorithm of the parameters estimation. On the first step we estimate parameters of a marginal distribution of the features. On the second step we estimate copula parameters.

Keywords: *ordinal scales, copula, Sklar theorem, integral indicators, expert estimations.*

Введение

Рассматривается задача построения интегрального индикатора в ранговых шкалах. Интегральный индикатор — это число, поставленное в соответствие объекту, и рассматриваемое как оценка его качества. Интегральными индикаторами называется вектор оценок, поставленный в соответствие набору объектов.

Ранее в работах [1, 2] были описаны процедуры построения интегральных индикаторов с использованием описаний объектов в линейных шкалах. При этом интегральный индикатор являлся уточнением оценки, заданной экспертом в линейной или ранговой шкале. В данной работе рассматривается ранговое описание объектов.

Описаниями объектов являются критерии, выбранные экспертами. Для построения интегрального индикатора оценивается совместная вероятность распределения критериев. Для решения этой задачи используются копулы [3, 4] — функции, являющиеся многомерными параметрическими функциями распределения равномерно распределенных случайных величин. Для оценки параметров распределения предлагается итеративный алгоритм: на первом шаге оцениваются параметры одномерных распределений критериев, на втором шаге оцениваются параметры копул.

Научный руководитель В.В. Стрижов
Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00709.

Использование копул для оценки распределений помогает справиться с проблемой ранговости критериев. Для оценки параметра копулы необходимо знать только ранговые соотношения между величинами, а не их абсолютные значения.

Предлагается алгоритм выбора наиболее информативных критериев, использующий параметры копулы в качестве показателя информативности. Эти параметры обладают свойством монотонности: чем больше параметр копулы, тем больше ранговая связь между двумя случайными величинами. Предлагается выбирать те критерии, которые имеют наибольшую ранговую связь с экспертными оценками интегральных индикаторов.

В качестве прикладной задачи рассматривается задача определения статуса угрожаемых видов животных, входящих в список Красной книги РФ [5, 7]. В Красной книге РФ принята следующая категоризация редкости видов (таксонов) по степени угрозы их исчезновения. Имеется шесть различных категорий статуса (интегральных индикаторов) таксонов: 0 — вероятно исчезнувшие, 1 — находящиеся под угрозой исчезновения, 2 — сокращающиеся в численности, 3 — редкие, 4 — неопределенные по статусу, 5 — восстанавливаемые и восстанавливающиеся. Эта категоризация является монотонной: интегральные индикаторы ранжированы по возрастанию биологического разнообразия.

Каждый таксон описан набором признаков, отражающих его состояние. Эксперт, владеющий информацией о таксоне, выставляет оценку для каждого признака в ранговой шкале. Таким образом, задана матрица «объект-признак», состоящая из описаний таксонов и вектор меток классов таксонов. Требуется построить модель, восстанавливающую интегральный индикатор таксона из Красной книги РФ по его описанию.

Задача ревизии Красной книги РФ и построения модели вычисления интегрального индикатора является актуальной из-за постоянного пополнения книги новыми записями о таксонах.

Постановка задачи

Пусть X — множество объектов, Y — конечное множество меток классов. Множество $X \times Y$ является вероятностным пространством с совместной функцией распределения $P(x, y)$.

Задано множество $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}, \quad i \in \mathcal{I} = \{1, \ldots, m\}$, которое является выборкой пар $(\mathbf{x}_i, y_i)$. Объект $\mathbf{x}_i \in X$— таксон, $y_i \in Y$ — метка класса, соответствующая этому таксону.

Описание объекта $x_i = [\boldsymbol{\chi}^1, \ldots, \boldsymbol{\chi}^j, \ldots, \boldsymbol{\chi}^n]^\intercal, \quad j \in \mathcal{J} = \{1, \ldots, n\}$ — это набор экспертных оценок признаков. Оценки объектов по признакам выставлены в ранговых шкалах. Каждый признак χ_j имеет собственную ранговую шкалу $\mathbb{L}_j$, состоящую из k_j упорядоченных элементов $\mathbb{L}_j = \{1 \prec 2 \prec \cdots \prec k_j\}$. Значение класса y также принадлежит упорядоченному множеству $\mathbb{L}_0 = \{1 \prec 2 \prec \cdots \prec k_0\}$.

Решается задача классификации объектов. Для этого предлагается найти отображение $a : X \to Y$, минимизирующее функционал среднего риска. Минимум среднего риска достигается алгоритмом

$$a(x) = \arg\max_{y \in Y} P(y|\mathbf{x}_i),$$

где $P(y|\mathbf{x}_i)$ — апостериорная вероятность класса y для объекта $\mathbf{x}$. Эта вероятность является условной по $\mathbf{x}$. Для оценки апостериорной вероятности $P(y|\mathbf{x}_i)$ будем использовать копулы.

Свойства копул, используемые для оценки условной вероятности

Определение 1. *Функция* $C : [0,1]^d \to [0,1]$ *называется копулой размерности* d*, если выполняются следующие условия:*

$$C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0,$$

$$C(1, \ldots, 1, u, 1, \ldots, 1) = u,$$

$$B = \prod_{i=1}^{d} [a_i, b_i] \subseteq [0,1]^d : \quad \int_B dC(u) \geqslant 0.$$

Выполнение этих свойств означает, что функция C является функцией распределения многомерной случайной величины $[u_1, \ldots, u_d]^\intercal$, такой, что одномерное распределение каждого из u_i равномерно на интервале $[0,1]$.

Важным фактом, позволяющим применять копулы для построения регрессионных моделей, является следующая теорема.

Теорема 1. *Многомерная функция распределения случайной величины:*

$$H(x_1, \ldots, x_d) = P[X_1 \leqslant x_1, \ldots, X_d \leqslant x_d]$$

случайного вектора $(X_1, \ldots, X_d)$ *с одномерными функциями распределения*

$$F_i(x) = P[X_i \leqslant x_i]$$

может быть записана в виде:

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

Таким образом, для оценивания совместного распределения H случайных величин $X_1, \ldots, X_d$ достаточно оценить их одномерные распределения $F_i(x_i)$ и функцию копулы, связывающую эти случайные величины.

Следующая теорема утверждает, что функция копулы не изменяется при действии на случайные величины любых монотонных преобразований.

Теорема 2. *Пусть* X*,* Y *— две случайные величины с совместной функцией распределения* $H(x, y)$*. Пусть также* φ, ψ*— две монотонных функции, преобразующие случайные величины* X *и* Y *в*

$$Z = \varphi(X), \quad T = \psi(Y)$$

с совместной функцией распределения $H'(Z, T)$*. Тогда копула, связывающая случайные величины* Z *и* T*:*

$$C'\big(F'(z), G'(t)\big) = H'(z, t) = C(F'(z), G'(t)),$$

то есть,

$$C' = C.$$

Таким образом, чтобы оценить функцию копулы, описывающую связь между случайными величинами $X_1, \ldots, X_d$, достаточно знать только ранговые соотношения этих случайных величин. Абсолютные значения величин $X_1, \ldots, X_d$ используются только при оценивании их одномерных распределений.

Для решения задачи классификации таксонов необходимо знать апостериорную вероятность (2). Эта вероятность выражается через частную производную функции копулы C, о чем утверждает следующая теорема.

Теорема 3. *Пусть X, Y — две случайные величины с одномерными функциями распределения $F(X)$, $G(Y)$. Тогда условная вероятность $P(Y \leqslant y|X = x)$ равна частной производной копулы:*

$$P(Y \leqslant y|X = x) = \frac{\partial}{\partial v} C(u, v)|_{(G(y),F(x))},$$

взятой в точке

$$u = G(y), \quad v = F(X).$$

В нашей задаче имеется n случайных величин, соответствующих признакам, и случайная величина Y.

Для оценки условной вероятности необходимо ввести некоторые дополнительные обозначения.

Имеется набор объектов $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$. Каждый объект описывается n признаками. Обозначим одномерные функции распределения y и всех компонент многомерной случайной величины $\mathbf{x}$:

$$G^0_Y(y), G^1_{X^1}(x^1), \ldots, G^n_{X^n}(x^n).$$

Обозначим совместные функции распределения упорядоченных поднаборов - векторов $\mathbf{x}^k = (x^1, \ldots, x^k)$ размерности от 1 до n:

$$F^k_{\mathbf{X}^k}(\mathbf{x}^k), \quad \mathbf{x}^k = (x^1, \ldots, x^k), \quad k = 1, \ldots, n.$$

Для нахождения условной вероятности $P(Y \leqslant y|\mathbf{x}_i)$, воспользуемся частной производной копулы $C(u, v)$ по переменной u:

$$P(Y \leqslant y|\mathbf{x}_i) = \frac{\partial}{\partial u} C(u, v)|_{F^n_{\mathbf{X}^n}(\mathbf{x}^n_i),G_Y(y)},$$

взятой в точке

$$u = F^n_{\mathbf{X}^n}(\mathbf{x}^n_i), \quad v = G_Y(y).$$

Неизвестной в этой формуле является функция совместного распределения $F^n_{\mathbf{X}^n}$. Чтобы найти эту функцию, воспользуемся теоремой 3:

$$F^n_{\mathbf{X}^n}(\mathbf{x}^n) = C^{n-1}(u, v)|_{F^{n-1}_{\mathbf{X}^{n-1}}(\mathbf{x}^{n-1}),G^n_{X_n}(x_n)},$$

$$\ldots,$$

$$F^i_{\mathbf{X}^i}(\mathbf{x}^i) = C^{i-1}(u, v)|_{F^{i-1}_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}),G^i_{X_i}(x_i)},$$

$$\ldots,$$

$$F^2_{X^1,X^2}(x^1, x^2) = C^1(u, v)|_{G^1_{X^1}(x^1),G^2_{X^2}(x^2)}.$$

Таким образом, чтобы оценить апостериорную вероятность $P(Y \leqslant y|\mathbf{x}_i)$, необходимо оценить все $n + 1$ одномерные распределения y и компонент случайного вектора $\mathbf{x}$, а также n копул $C, C^1, \ldots, C^{n-1}$.

Копулы, используемые при построении интегрального индикатора

Для решения задачи (2) предлагается использовать Архимедовскую копулу:

Определение 2. *Копула $C(u_1, \ldots, u_d)$ называется архимедовской, если для нее выполнены следующие условия:*

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)),$$

где функция ψ называется генератором, и для нее должны быть выполнено:

$$(-1)^k \psi^{(k)}(x) \geqslant 0$$

для всех $x \geqslant 0$ и $k = 0, 1, \ldots, d-2$. А также, функция

$$(-1)^{d-2} \psi^{d-2}(x)$$

должны быть невозрастающей и выпуклой.

Будем использовать частные случаи Архимедовской копулы, задаваемые следующими функциями-генераторами:

копула Клейтона,

$$\psi(t) = (1 + \theta t)^{-\frac{1}{\theta}}, \quad \theta \in \Theta = (0, \infty)$$

и копула Гумбеля,

$$\psi(t) = \exp(-t^{\frac{1}{\theta}}), \theta \in \Theta = [1, \infty).$$

Отметим, что эти семейства копул зависят только от одного параметра θ, что значительно упрощает задачу в вычислительном смысле.

В случае копулы Гумбеля, частная производная имеет следующий вид:

$$\frac{\partial}{\partial u} C(u, v) = \left(\frac{\ln u}{\ln C}\right)^{\theta - 1} \frac{C}{u}.$$

Оценка параметров копулы

Как было сказано выше, для оценки параметра $\theta \in \Theta$ копулы используются не сами случайные величины X, Y, а последовательности рангов этих величин. Выборкам X и Y соответствуют последовательности рангов:

$R_x = (R_{x_1}, \ldots, R_{x_m})$, где $R_{x_i} -$ ранг $i-$го объекта в вариационном ряду выборки X,

$R_y = (R_{y_1}, \ldots, R_{y_n})$, где $R_{y_i} -$ ранг $i-$го объекта в вариационном ряду выборки Y.

Отметим, что наиболее часто используемым методом оценки параметров распределения является метод максимизации правдоподобия, который в случае копул записывается следующим образом:

$$L(\theta) = \sum_{i=1}^{m} \log\Big(c_\theta\big(F(X_i), G(Y_i)\big)\Big),$$

$$c_\theta(u, v) = \frac{\partial^2}{\partial u \partial v} C_\theta(u, v).$$

Вместо значений функций одномерных распределений $F(X_i), G(Y_i)$ можно подставить их эмпирические значения, получив таким образом функцию псевдоправдоподобия [6]:

$$L'(\theta) = \sum_{i=1}^{m} \left(\log c_\theta \left(\frac{R_i}{m+1}, \frac{S_i}{m+1} \right) \right).$$

Заметим, что функция L' зависит только от самой копулы C_θ, то есть, в нашем случае, только от параметра θ, и максимизация этой функции не представляет собой большой вычислительной сложности.

Благодаря этому способу, задача оценки распределений $F^i_{\mathbf{X}^i}(\mathbf{x}^i)$ распадается на два независимых этапа: оценка параметра θ_i копул C^i путем максимизации псевдоправдоподобия и оценка параметров одномерных распределений $G^0_Y(y), G^1_{X^1}(x^1), \ldots, G^n_{X^n}(x^n)$ с помощью метода максимума правдоподобия.

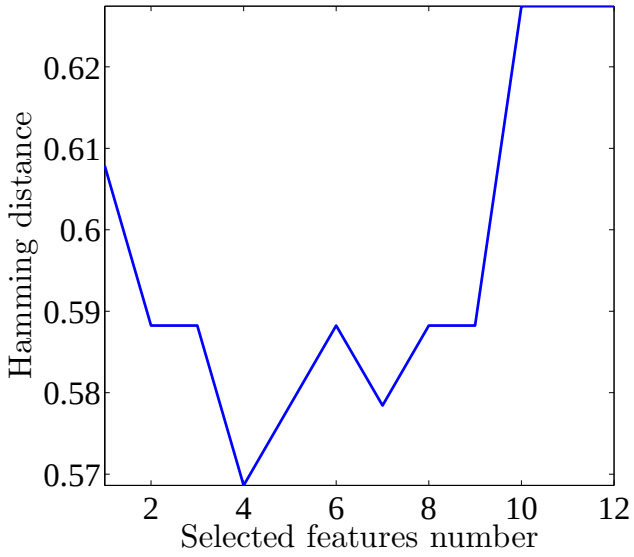


Рис. 1: Зависимость ошибки классификации от количества выбранных признаков

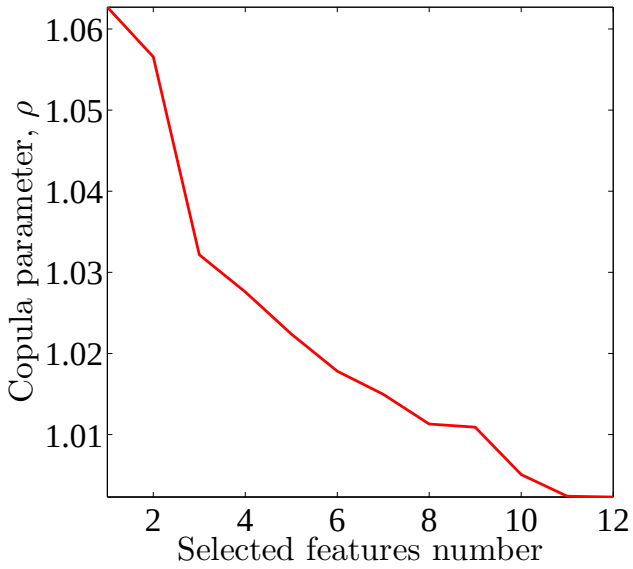


Рис. 2: Зависимость параметра копулы от количества выбранных признаков

Алгоритм оценки апостериорного распределения

Приведем подробный алгоритм оценки распределений $F^i_{\mathbf{X}^i}(\mathbf{x}^i)$. Как было сказано выше, необходимо оценить $n+1$ одномерное распределение $G^0_Y(y), G^1_{X^1}(x^1), \ldots, G^n_{X^n}(x^n)$ и n функций копулы $C, C^1, \ldots, C^n$.

1. Оцениваются одномерные распределения $G^1_{X^1}(x^1)$, $G^2_{X^2}(x^2)$. Все функции $G^i_{X^i}(x^i)$ будем искать в классе бета-распределений. То есть, распределение случайной величины X задается плотностью вероятности g_X, имеющей вид:

$$\begin{cases} g_X(x) & = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \\ B(\alpha,\beta) & = \int\limits_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx. \end{cases}$$

Параметры α и β для этого распределения оцениваются методом моментов. Для этого численно решается система уравнений:

$$\begin{cases} E(X) & = \frac{\alpha}{\alpha+\beta}, \\ D(X) & = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{cases}$$

2. Оценим копулу $C^1(u,v)$, связывающую переменные x^1 и x^2, максимизируя функцию псевдоправдоподобия:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L'_{12}(\theta) = \sum_{i=1}^{m} \left(\log c_\theta \left(\frac{R_{x_1}}{m+1}, \frac{R_{x_2}}{m+1} \right) \right).$$

3. Оценив одномерные распределения $G^1_{X^1}(x^1)$, $G^2_{X^2}(x^2)$ и копулу $C^1(u,v)$, получаем оценку функции совместного распределения $F^2_{X^1,X^2}(x^1,x^2)$. Повторяем шаги 1-2, каждый раз прибавляя по одному новому признаку x^i и оценивая на шаге 3 функцию $F^i_{\mathbf{X}^i}(\mathbf{x}^i)$.

4. повторив n раз шаги 1-2, получим функцию совместного распределения всех признаков $F^n_{\mathbf{X}^n}$. На последнем шаге оценим функцию распределения $G^0_Y(y)$, копулу $C(u,v)$, связывающую Y и X, и найдем $\hat{y}$, доставляющий максимум апостериорной вероятности:

$$\hat{y} = \arg\max_y P(Y \leqslant y|\mathbf{x}_i) = \arg\max_y \frac{\partial}{\partial u} C(u,v)|_{F^n_{\mathbf{X}^n}(\mathbf{x}^n_i),G_Y(y)},$$

взятой в точке

$$u = F^n_{\mathbf{X}^n}(\mathbf{x}^n_i), \quad v = G_Y(y),$$

где

$$F^i_{\mathbf{X}^n}(\mathbf{x}^i_i) = C^{i-1}(u,v)|_{F^{i-1}_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}),G^i_{X_i}(x_i)},$$

взятой в точке

$$u = F^{i-1}_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}), \quad v = G^i_{X_i}(x_i)$$

для всех

$$i = 2, \ldots, n.$$

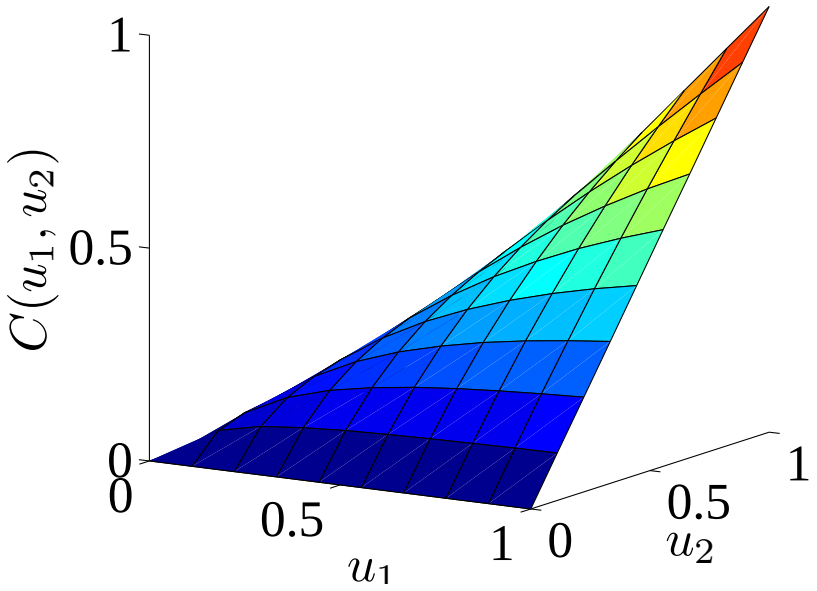


Рис. 3: Копула Клейтона

Выбор признаков

Так как число объектов в данной задаче, определенное составом Красной книги РФ, сопоставимо с числом признаков, необходимо выбрать наиболее информативные признаки. Множество индексов признаков, включенных в функцию вероятности 2, назовем активным набором и обозначим $\mathcal{A} \subseteq \mathcal{J}$.

Для того, чтобы выбрать наиболее информативные признаки, предлагается использовать следующий эвристический алгоритм. Информационными будем читать те признаки,

которые имеют наибольшую ранговую связь со случайной величиной Y. Чтобы понять, какие признаки имеют наибольшую связь, рассмотрим некоторые свойства копул о ранговой связи.

Утверждение 1. *Случайные величины X и Y являются независимыми тогда и только тогда, когда*

$$C(u,v) = uv, \quad u,v \in [0,1],$$

где

$$C(F(x),G(y)) = H(x,y),$$

где $H(x,y)$ — совместная функция распределения случайных величин X и Y.

Утверждение 2. *Границы Фреше для копулы:*

$$W(u,v) \leqslant C(u,v) \leqslant M(u,v), \quad u,v \in [0,1],$$

где

$$W(u,v) = \max(0, u+v-1)$$

— минимальная копула,

$$M(u,v) = \min(u,v)$$

— максимальная копула.

Причем, если $C(u,v) = W(u,v)$, то Y— монотонно убывающая функция X, если $C(u,v) = M(u,v)$, то Y— монотонно возрастающая функция X.

Для примера, рассмотрим копулу Гумбеля (2):

$$C_\theta(u,v) = \exp\left[\left((-\log(u))^\theta + (-\log(v))^\theta\right)^{\frac{1}{\theta}}\right] \quad \theta \geqslant 1.$$

При стремлении параметра копулы $\theta \to 1$, $C_\theta(u,v) \to uv$, то есть, случайные величины являются независимыми. При стремлении параметра $\theta \to \infty$, ранговая связь между случайными величинами возрастает. Таким образом, ранговая связь изменяется монотонно при варьировании параметра копулы. Для решения задачи отбора признаков будем отбирать те из них, для которых параметр копулы со случайной величиной Y является наибольшим.

Исходя из этого рассуждения, предлагается следующий алгоритм.

1. Примем пустое множество активных признаков

$$\mathcal{A} = \varnothing.$$

2. Для всех $j = 1,\ldots,n$ вычислим параметры θ_j для копул $C_{\theta_j}(F_i(x^j), G(y))$ и включим в набор

$$\mathcal{A} = \mathcal{A} \cup \{k\}$$

тот признак k, для которого

$$k = \arg\max_{j \in \mathcal{J}} \theta_j.$$

Обозначим множество оставшихся признаков

$$\mathcal{J}' = \mathcal{J} \setminus \mathcal{A}.$$

3. Для всех признаков $j \in \mathcal{A}$ и всех $k_1, \ldots, k_{\mathcal{A}}$ вычислим параметры θ_j для копул

$$C_{\theta_j}(F_i(x_i), H_{k_1,\ldots,k_{\mathcal{A}}}(x^{k_1}, \ldots, x^{k_{\mathcal{A}}}))$$

и включим в набор

$$\mathcal{A} = \mathcal{A} \cup \{k\}$$

тот признак k, для которого

$$k = \arg\max_{j \in \mathcal{J}'} \theta_j.$$

4. Будем повторять шаг 3, пока значение ошибки на контрольной выборке не стабилизируется.

Вычислительный эксперимент

Работа алгоритма иллюстрируется данными из Красной Книги РФ. Экспертами заполнена таблица данных для 29 различных объектов. Каждый объект описывается 102 признаками.

На рис. 1 показана зависимость ошибки классификации от количества выбранных признаков. Оптимальное значение достигается при $|\mathcal{A}| = 4$. В исходной таблице данных эти признаки индексированы номерами 22, 24, 23 и 20.

На рис. 2 показана зависимость параметра копулы от количества выбранных признаков. Видно, что значение параметра монотонно убывает с ростом количества признаков.

Литература

[1] Стрижов В. В. Уточнение экспертных оценок с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов. 2006, Т. 72(7). С. 59–64.

[2] Strijov V., Granic G., Juric J., Jelavic B., Maricic S.A. Integral indicator of ecological impact of the Croatian thermal power plants // Energy, 2011. Vol. 36(7). Pp. 4144–4149.

[3] Roger B. Nelsen An Introduction to Copulas // Springer, 1998

[4] Edward W. Frees and Emiliano A. Valdez Understanding relationships using copulas // North american actuarial journal, 2012. Vol. 2. Pp. 104-141.

[5] Красная книга Российской Федерации. М.: Институт проблем экологии и эволюции имени А. Н. Северцова РАН / Под ред. В. И. Данилов-Данильян и др. http://www.sevin.ru/redbook/ 31.07.2012.

[6] Christian Genest and Anne-Catherine Favre *Everything you always wanted to know about copula modeling but were afraid to ask* // Journal of Hydrologic Engineering, 2007. P. 347–368

[7] Красная книга Российской Федерации (животные). М: АСТ Астрель, 2001.

[8] Стрижов В. В. Уточнение экспертных оценок, выставленных в ранговых шкалах, с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов. 2011, Т. 77(7). С. 72–78.

[9] Литвак Б. Г. Экспертная информация: Методы получения и анализа. М.: Радио и связь, 1982. С. 69–88.

[10] Орлов А. И. Организационно-экономическое моделирование: часть 2. Экспертные оценки. М: МГТУ им. Н. Э. Баумана, 2011. 486 с.

[11] Boyd S. and Vandenberghe L. Convex Optimization // Cambridge University Press. 2004.

Вероятностная модель одноклассовой классификации*

Бурмистров М. О., Сандуляну Л. Н.
burmisha@gmail.com, liubov.sanduleanu@gmail.com
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Решается задача одноклассовой классификации электронных писем на предмет наличия в них спама. В работе вводится квазивероятностная модель для классической эмпирической постановки задачи одноклассовой классификации и производится сведение классического подхода к новой модели. Построенные методы классификации проверяются вычислительными экспериментами на модельных и реальных данных.

Ключевые слова: *одноклассовая классификация, вероятностная модель, байесовский подход, ядерные функции.*

Probabilistic model for one-class classification problem*

Burmistrov M. O., Sanduleanu L. N.
Moscow Institute of Physics and Technology

One-class classification methods are used to test e-mails for spam. Quasi-probabilistic model is introduced for traditional empirical approach to problem. The old model is shown to be a reduction of the new one. Built approaches to classification are numerically tested on model and real data.

Keywords: *one-class classification, probabilistic model, Bayesian approach, kernel functions.*

Введение

С широким развитием сети Интернет и ее проникновением в большую часть всех сфер жизни у людей появилась возможность свободно обмениваться информацией и получать доступ к разнообразным ресурсам. Одним из наиболее распространенных способов общения людей через Интернет является использование электронной почты. В силу большой открытости этого канала связи с точки зрения возможности передачи любого сообщения произвольному пользователю он активно используется мошенниками, злоумышленниками и распространителями рекламных материалов. При этом создается не только повышенная нагрузка на техническую инфраструктуру, но и тратится время людей, которым приходится отделять полезную информацию от всей остальной. Поэтому задача автоматизации фильтрации электронной почты будет оставаться актуальной в течение всего времени ее существования.

Задача фильтрации спама уже решалась различными методами [1, 2], однако они в большой степени являлись эвристическими и не имели под собой четкой вероятностной модели. Также проблемой является корректное составление обучающей выборки. Дело в том, что спам-письма зачастую шаблонны и имеют много общего в своей структуре, к тому же они широко доступны. Составить же обучающую выборку, содержащую письма, полезные для пользователей, гораздо сложнее по следующим причинам:

— меньшая доступность,
— высокая разнородность,

Научный руководитель О. В. Красоткина

— большое число шаблонных писем (разнообразные уведомления от сервисов).

По этим причинам предлагается использовать методы одноклассовой классификации [3, 4], чтобы отказаться от требования к обучающей выборке содержать достаточно широкое множество разнообразных представителей обоих классов.

В работе предложена квазивероятностная постановка задачи одноклассовой классификации. Такой подход позволяет уточнить область применимости построенной модели и предъявляемые требования к данным. На основе полученной вероятностной постановки задачи строится новая вероятностная модель порождения объектов, в ходе оптимизации которой происходит построение классификатора.

Полученные методы построения одноклассовых классификаторов применяются к модельным и реальным данным.

Байесовская постановка задачи

Объектом исследования является множество электронных сообщений, характеризуемых некоторым набором признаков. Рассмотрим одноклассовую классификацию объектов генеральной совокупности Ω. Пусть каждый объект $\omega \in \Omega$ представлен точкой в линейном пространстве признаков $\mathbf{x}(\omega) = (x^1(\omega), \ldots, x^n(\omega)) \in \mathbb{R}^n$. При этом мы изучаем лишь объекты одного класса, поэтому меткой класса объект существенно не обладает. Тем не менее нашей задачей будет построение классификатора, который будет давать ответ 1, если предъявленный объект лежит в множестве, и 0 — иначе.

В работе [3] предлагается строить сферический пороговый классификатор вида $[z \leqslant 0]$, где $z(\mathbf{x}, \mathbf{a}, R) = \|\mathbf{x} - \mathbf{a}\| - R$ без вероятностного обоснования такого подхода. При этом в области $z(\mathbf{x}, \mathbf{a}, R) \geqslant 0$ значение величины $\|\mathbf{x} - \mathbf{a}\|^2 - R^2$ несет смысл отступа ξ, а для объектов внутри шара отступ полагается равным 0. Для подбора значений $\mathbf{a}, R$ решается задача

$$F(R, \mathbf{a}, \boldsymbol{\xi}) = R^2 + C \sum_i \xi_i \to \min_{\mathbf{a}, R, \boldsymbol{\xi}}, \tag{1}$$

при этом здесь и далее мы полагаем, что суммирование по индексу i (а в дальнейшем и j) означает суммирование по всем объектам обучающей выборки.

Здесь величина C задает баланс между минимальным объемом шара и наименьшим числом объектов обучающей выборки вне сферы. Пример описания объектов шаром приведен на рис. 1.

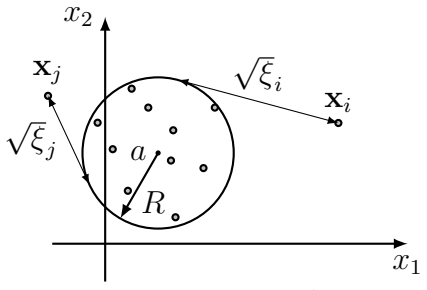


Рис. 1: Пример описания объектов шаром

Будем придерживаться вероятностной модели распределения объектов генеральной совокупности. Параметрическое семейство условных плотностей распределения в призна-

ковом пространстве имеет вид

$$\varphi\left(\mathbf{x}|\mathbf{a},R;c\right)\propto\begin{cases}1, & z(\mathbf{x},\mathbf{a},R)<0,\\ e^{-c\left(\|\mathbf{x}-\mathbf{a}\|^2-R^2\right)}, & z(\mathbf{x},\mathbf{a},R)\geqslant 0.\end{cases}\tag{2}$$

Здесь величина c является гиперпараметром. График данной функции плотности изображен на рис. 2.

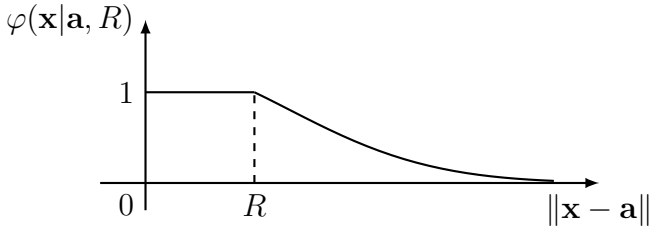


Рис. 2: Значение плотности распределения вдоль радиуса

Совместную плотность распределения случайной обучающей совокупности будем понимать как плотность распределения выборки независимых реализаций

$$\Phi(\mathbf{X}|\mathbf{a},R)=\prod_{j=1}^{N}\varphi(\mathbf{x}_j|\mathbf{a},R),$$

где $\mathbf{X}=\{\mathbf{x}\}_{j=1}^{N}$. Пусть, далее, выбрана априорная плотность совместного распределения вероятностей $\Psi(\mathbf{a},R)$ для параметров распределения $\varphi\left(\mathbf{x}|\mathbf{a},R;c\right)$. Тогда апостериорная плотность распределения параметров $\mathbf{a}$ и R относительно обучающей совокупности определяется формулой Байеса

$$p(\mathbf{a},R|\mathbf{X})=\frac{\Psi(\mathbf{a},R)\Phi(\mathbf{X}|\mathbf{a},R)}{\int\Psi(\mathbf{a}',R')\Phi(\mathbf{X}|\mathbf{a}',R')d\mathbf{a}'dR'}.\tag{3}$$

Из принципа максимума плотности апостериорного распределения в пространстве параметров модели генеральной совокупности получим байесовское правило обучения

$$\left(\hat{\mathbf{a}},\hat{R}|\mathbf{X}\right)=\arg\max_{\mathbf{a},R}p(\mathbf{a},R|\mathbf{X})\tag{4}$$

Поскольку знаменатель в выражении (3) не зависит от целевых переменных

$$p(\mathbf{a},R|\mathbf{X})\propto\Psi(\mathbf{a},R)\Phi(\mathbf{X}|\mathbf{a},R)=\Psi(\mathbf{a},R)\prod_{j=1}^{N}\varphi(\mathbf{x}_j|\mathbf{a},R),$$

то в задаче максимизации (4) достаточно рассматривать только числитель

$$\left(\hat{\mathbf{a}},\hat{R}|\mathbf{X}\right)=\arg\max_{\mathbf{a},R}p(\mathbf{a},R|\mathbf{X})=\arg\max_{\mathbf{a},R}\left(\ln\Psi(\mathbf{a},R)+\sum_{j=1}^{N}\ln\varphi(\mathbf{x}_j|\mathbf{a},R)\right).$$

Теперь покажем, что задача в такой постановке обобщает задачу (1). Положим, что априорное распределение параметров $\Psi(\mathbf{a},R)$ обладает следующими свойствами:

— $\mathbf{a}$ и R — случайные независимые величины,
— $|R|$ — нормально распределенная случайная величина с нулевым математическим ожиданием и дисперсией σ^2,
— $\mathbf{a}$ равномерно распределено по всему пространству $\mathbb{R}^n$ (такое распределение будет несобственным [5]).

Тогда совместное распределение параметров также будет несобственным

$$\Psi(\mathbf{a}, R) \propto e^{-\frac{1}{2\sigma^2}R^2}.$$

Подставим это выражение и функцию распределения из (2)

$$\begin{aligned}\ln p(\mathbf{a}, R|\mathbf{X}) &= \ln \Psi(\mathbf{a}, R) + \sum_{j=1}^{N} \ln \varphi(\mathbf{x}_j|\mathbf{a}, R) = \\ &= -\frac{R^2}{2\sigma^2} + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|\leqslant R} \ln 1 + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \ln e^{-c\left(\|\mathbf{x}-\mathbf{a}\|^2-R^2\right)} = \\ &= -\frac{R^2}{2\sigma^2} - \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} c\left(\|\mathbf{x}_i-\mathbf{a}\|^2 - R^2\right) = \\ &= -\frac{1}{2\sigma^2}\left(R^2 + 2\sigma^2 c \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \left(\|\mathbf{x}_i-\mathbf{a}\|^2 - R^2\right)\right) \to \max_{\mathbf{a},R}. \end{aligned} \tag{5}$$

Очевидно, задачи (5) и (1) эквивалентны при $C = 2\sigma^2 c$.

Решение оптимизационной задачи

Итак, для нахождения значений $\mathbf{a}$ и R необходимо решить следующую задачу

$$\begin{cases} R^2 + C\sum\limits_i \xi_i \to \min\limits_{\mathbf{a},R,\boldsymbol{\xi}}, \\ \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant R^2 + \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, N. \end{cases} \tag{6}$$

Функция Лагранжа этой задачи имеет вид

$$\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = R^2 + C\sum_i \xi_i - \sum_i \gamma_i\xi_i - \sum_i \alpha_i\left(R^2 + \xi_i - (\mathbf{x}_i^\intercal\mathbf{x}_i - 2\mathbf{a}^\intercal\mathbf{x}_i + \mathbf{a}^\intercal\mathbf{a})\right),$$

где $\alpha_i \geqslant 0$ и $\gamma_i \geqslant 0$ — множители Лагранжа. Необходимым условием минимума является равенство нулю частных производных функции Лагранжа по всем переменным

$$\begin{aligned} &\tfrac{\partial\mathcal{L}}{\partial R} = 0: \quad \sum_i \alpha_i = 1 \text{ (случай } R = 0 \text{ рассмотрим отдельно,)} \\ &\tfrac{\partial\mathcal{L}}{\partial \mathbf{a}} = 0: \quad \mathbf{a} = \frac{\sum\limits_i \alpha_i\mathbf{x}_i}{\sum\limits_i \alpha_i} = \sum_i \alpha_i\mathbf{x}_i, \\ &\tfrac{\partial\mathcal{L}}{\partial \xi_i} = 0: \quad \gamma_i = C - \alpha_i, \quad i = 1, \ldots, N. \end{aligned} \tag{7}$$

Из последнего уравнения получаем, что $\alpha_i = C - \gamma_i$. Таким образом, мы получаем новые ограничения на α_i

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N.$$

Если это ограничение выполнено, то мы можем вычислить γ_i по формуле $\gamma_i = C - \alpha_i$, и при этом автоматически будет выполнено условие $\gamma_i \geqslant 0$.

Тогда для функции Лагранжа получим выражение

$$\begin{aligned}\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= R^2 - \sum_i \alpha_i R^2 + C\sum_i \xi_i - \sum_i \alpha_i \xi_i + \\ &+ \sum_i \alpha_i \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - 2\sum_i \alpha_i \mathbf{a}^{\mathsf{T}}\mathbf{x}_i + \sum_i \alpha_i \mathbf{a}^{\mathsf{T}}\mathbf{a} - \sum_i \gamma_i\xi_i = \\ &= R^{\mathsf{T}}R^{\mathsf{T}}\left(1 - \sum_i \alpha_i\right) + \sum_i \xi_i\,(C - \alpha_i - \gamma_i) + \\ &+ \sum_i \alpha_i \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - 2\sum_i \alpha_i \sum_j \alpha_j \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_i + \sum_{i,j} \alpha_i\alpha_j \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_i = \\ &= \sum_i \alpha_i \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - \sum_{i,j} \alpha_i\alpha_j \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_i \to \max_{\boldsymbol{\alpha}}.\end{aligned}$$

Полученное выражение является квадратичной формой. Тогда его максимум находится по известным алгоритмам решения задач квадратичного программирования. По оптимальным значениям $\boldsymbol{\alpha}$ мы сможем найти оптимальное значение центра гипершара $\mathbf{a}$ и отступов $\boldsymbol{\xi}$, используя соотношения (7).

Для каждого объекта $\mathbf{x}_i$ оптимальное значение α_i (или же $\gamma_i = C - \alpha_i$) задает тип принадлежности объекта построенному гипершару:

— $\alpha_i = 0 \Rightarrow$ объект $\mathbf{x}_i$ лежит внутри гипершара, имеет нулевой отступ;
— $0 < \alpha_i < C \Rightarrow$ объект $\mathbf{x}_i$ лежит на границе гипершара, имеет нулевой отступ;
— $\alpha_i = C \Rightarrow$ объект $\mathbf{x}_i$ лежит вне гипершара, имеет ненулевой отступ.

Радиус R определяется как расстояние от центра гипершара $\mathbf{a}$ до опорных векторов, лежащих на границе гипершара.

Если же $R = 0$, то задача (6) имеет вид

$$\begin{cases} C\sum\limits_i \xi_i \to \min\limits_{\mathbf{a},\boldsymbol{\xi}}, \\ \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, N. \end{cases} \tag{8}$$

т.е.

$$C\sum_i \|\mathbf{x}_i - \mathbf{a}\|^2 \to \min_{\mathbf{a}}, \tag{9}$$

а эта задача соответствует методу наименьших квадратов. Тогда $\mathbf{a} = \frac{\sum_i \mathbf{x}_i}{N}$. При этом следует понимать, что значение $R = 0$ обнуляет обобщающую способность нашего классификатора, поэтому следует отказываться от такого решения, если есть выбор. Здесь же стоит отметить, что $R = 0$ обязательно, если $C < \frac{1}{N}$, где N — число объектов в обучающей выборке, поскольку в этом случае условия на $\boldsymbol{\alpha}$ несовместны.

Для возможности описания данных более гибкой формой, нежели сфера, в работе [3] предлагается использовать потенциальные функции [6]. Наиболее часто используемыми потенциальными функциями являются полиномиальная

$$K_p(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j)^p$$

и радиальная базисная функция Гаусса

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2s^2}\right).$$

Таким образом, чтобы получить улучшенную модель описания данных, необходимо заменить в функции Лагранжа операцию вычисления скалярного произведения двух векторов вычислением значения потенциальной функции двух аргументов.

Численный эксперимент

Для оценки качества работы алгоритма предлагается ввести метрику. Следуя работе [7], будем измерять качество одноклассовой класификации в терминах точности и полноты. В нашем случае точность (precision) — доля верно классифицированных объектов тестовой выборки среди всех объектов, отнесенных алгоритмом к единственному классу. Полнота (recall) — доля верно классифицированных объектов тестовой выборки среди всех объектов, принадлежащих к единственному классу. Более высокие значения точности и полноты соответствуют лучшему качеству классификации. В качестве агрегированного показателя, объединяющего точность P и полноту R используем F_1-меру [8]:

$$F_1 = \frac{2PR}{P+R}.$$

Для проведения вычислительного эксперимента сгенерируем $N = 400$ случайных точек $\{\mathbf{x}_i\}_{i=1}^N$ из распределения (2) при размерности пространства 2 (для наглядности), положив направления смещений случайными и придав параметрам значения $a = (1, 2)^\intercal$, $R = 3, c = 0{,}2$. После этого проведем $t\times q$-fold кросс-валидацию с $t = 10, q = 3$, скользящим контролем подбирая параметр C и вычисляя F_1-метрику при каждом его значении. При этом все, что лежит вне сферы, мы считаем не принадлежащим классу, а все, что внутри, — считаем. В результате получим следующую зависимость значения метрики от C (см. рис. 4). Из графика видно, что при $C \to 0$ обобщающая способность также стремится к нулю, поскольку практически отсутствует штраф за непопадание в класс при обучении. При этом большие штрафы заставляют необоснованно увеличивать сферу, снижая точность.

Пример работы алгоритма приведен на рис. 3 при параметре $C = 0{,}007$. Зеленым изображена граница истинного распределения, красным — построенного. Видно, что здесь C слишком мало и сфера получилось слишком маленькой.

Для проведения эксперимента на реальных данных были выбраны доступные в открытом доступе уже вычисленные признаки сообщений[1]. Здесь для обучения бралась небольшая часть спам-документов (200 из 1800). Сперва они линейно отображались в куб $[0, 1]^k$ ($k = 57$ — размерность пространства), а затем по ним строилась сфера в этом 57-мерном пространстве. Для контроля все остальные данные преобразовывались по тому же правилу (что не гарантирует их попадание в этот же куб), после чего проверялось попадание в построенную сферу и вычислялась F_1-метрика. Здесь в контроле уже участвуют объекты как из исследуемого класса (спам-сообщений), так и не из него, хотя обучение происходило только на объектах целевого класса. Результаты подбора параметра C изображены на рис. (4). Данные усреднены по 20 случайным выборкам по 200 объектов из 1800.

В обоих экспериментах отчетливо прослеживаются максимумы метрики, что свидетельствует о наличии в обеих задачах оптимального значения параметра C.

[1]UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Spambase

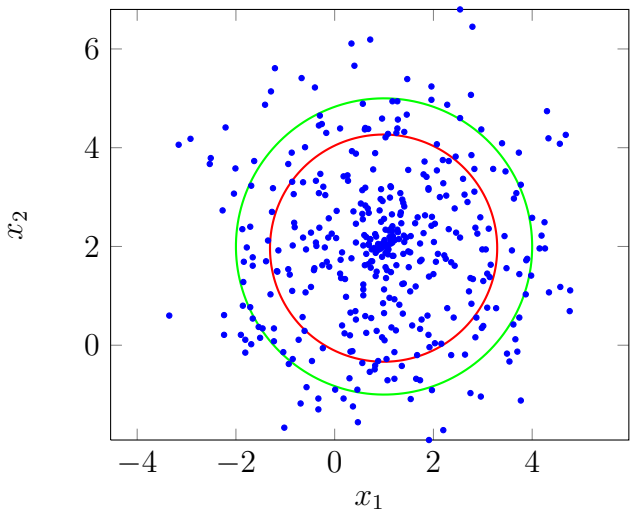


Рис. 3: Пример результата работы алгоритма при $C = 0{,}007$.

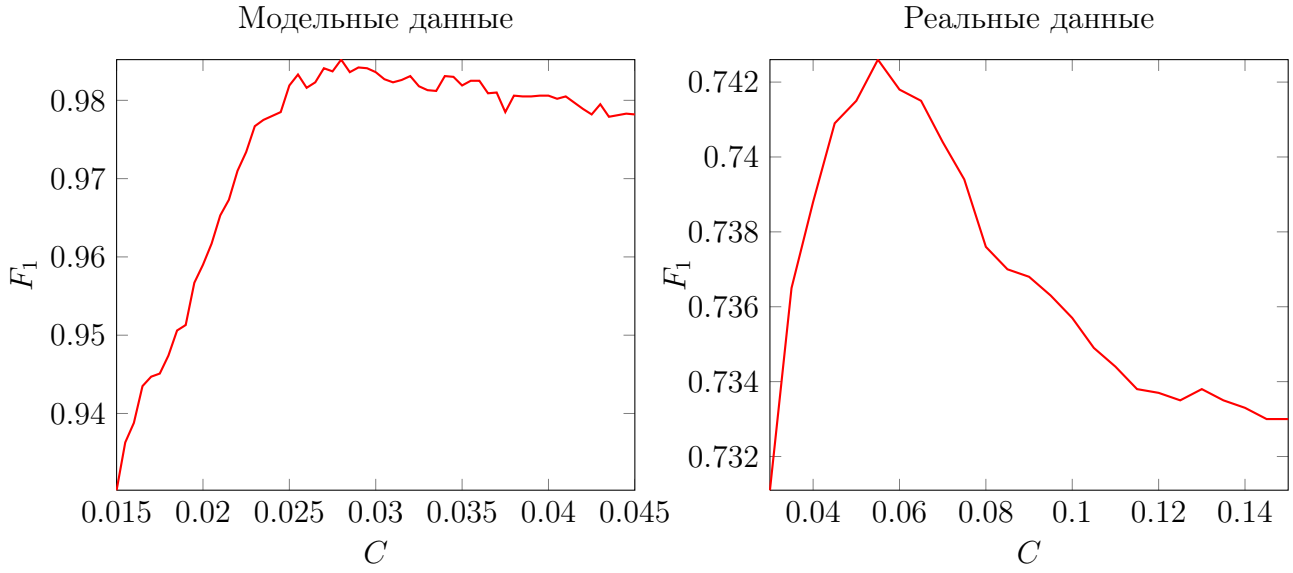


Рис. 4: Зависимость F_1-метрики от параметра регуляризации C

Заключение

В работе был предложен вероятностный подход к задаче одноклассовой классификации. Такой подход более удобен, чем классический подход, основанный на эвристических соображениях, и с теоретической, и с практической точек зрения, поскольку несет в себе ясную возможность модификаций. При этом классический подход является частным случаем предложенного алгоритма.

В работе построен алгоритм описания классов шарами, проведено его обобщения на случай ядерных функций. Проведены вычислительные эксперименты на модельных и реальных данных.

Литература

[1] Islam R., Chowdhury R. *Spam filtering using ML algorithms* // Universitetets Okonomiske Institute. IADIS International Conference on WWW/Internet. 2007.

[2] *Research of Spam Filtering system based on LSA and SHA* / Sun J. [et al.] // Advances in neural networks. ISNN. 2008.

[3] Tax D. *One-class classification; Concept-learning in the absence of counterexamples* // Ph.D thesis. 2001.

[4] Khan S., Madden G. *A Survey of Recent Trends in One Class Classiffcation*//College of Engineering and Informatics, National University of Ireland Galway. Ireland, 2006.

[5] Де Гроот М. *Оптимальные статистические решения* — М.: Мир, 1974.

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

[7] Романенко А. А. *Категоризация текстов на основе монотонного классификатора ближайшего соседа* // Выпускная валификационная работа бакалавра. 2012.

[8] van Rijsbergen C. J. *Information Retrieval* // Butterworth. 2nd ed. 1979.

Оценка плотности совместного распределения*

Мотренко А. П.
anastasia.motrenko@gmail.com
Московский физико-технический институт

В задачах классификации часто возникает ситуация, когда часть переменных распределена непрерывно, а часть — дискретно. Например, в логистической регрессии признаки непрерывны, а переменная отклика подчиняется распределению Бернулли. В работе описан способ оценки плотности совместного неоднородного распределения, включающего дискретные и непрерывные величины. Рассмотрен случай, когда вероятностные предположения о распределении случайных величин сделать не удается. В этом случае применяются методы ядерного сглаживания. В работе также приводится их сравнение с классическими методами теории вероятностей. Эксперимент проводится на реальных и синтетических данных.

Ключевые слова: *плотность совместного распределения, смешанное распределение, ядерное сглаживание, порождающие алгоритмы классификации.*

Joint probability density estimation*

A. P. Motrenko
Moscow Institute of Physics and Technology

When solving a classification problem one often has to deal with both discrete and continuous variables. for example, in the logistic regression independent variables are distributed continuously, while a target variable follows Bernoulli distribution. In this paper a method is resented that allows to estimate joint probability distribution which include discrete and continuous variables. A case when no probabilistic assumptions can be made is considered. The methods of nonparametric regression are used. Also a comparison to the classic methods of probability theory is presented. The experiment is conducted on the real and synthetic data.

Keywords: *joint distribution density, mixed distribution, nonparametric regression, generative classification algorithms..*

Введение

В задачах классификации требуется, по набору наблюдаемых величин, определить метку класса зависящей от них случайной величины. Алгоритмы классификации, включающие оценку плотности совместного распределения зависимых и независимых переменных, называются порождающими, так как с помощью восстановленной плотности совместного распределения можно породить пары зависимых и независимых переменных. Примерами порождающих алгоритмов являются наивный байесовский классификатор [1, 2], скрытые марковские цепи [3, 4].

В случае, когда наблюдаются реализации некоторой непрерывной случайной величины, а зависимая переменная подчиняется дискретному распределению, возникает задача оценки плотности смешанного совместного распределения, включающего дискретные и непрерывные случайные величины. При известных условных и маргинальных [5] плотностях рассматриваемых величин, плотность совместного распределения можно получить

Работа поддержана грантом РФФИ 12-07-31095.

аналитически, воспользовавшись определением условной вероятности. Такой способ называется факторизацией, он рассмотрен в работах [6, 7]. В работах [8, 9] рассматривается оценка плотности совместного распределения с помощью копул. В этом случае не делается предположений об условном распределении зависимой переменной при заданном наборе наблюдаемых переменных. Достаточно знать одномерные плотности распределения зависимой и независимых переменных.

В данной работе особое внимание уделяется случаю, когда сделать какие-либо предположения об условной зависимости распределений или о виде копулы не удается. В таком случае применяются методы непараметрического [10, 11] восстановления плотности. В данной работе методы ядерного сглаживания применяются для оценки совместного распределения дискретных и непрерывных случайных величин. Вычислительный эксперимент проводится на реальных и синтетических данных.

Факторизация

Рассмотрим вначале способ оценки плотности совместного распределения, основанный на разбиении ее на множители. Этот метод будет применяться в вычислительном эксперименте для получения эталонной оценки плотности распределения.

Функцию совместного распределения смешаной случайной величины $Z = (\mathbf{x}, y)$, где $y \in \{0, 1\}$ — дискретная случайная величина, $\mathbf{x} \in \mathbb{R}^n$ — вектор непрерывных независимых случайных величин, определим как

$$P(\mathbf{x}, y) = \sum_{t \leqslant y} \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} p(\mathbf{s}, t) \mathrm{d}\mathbf{s}, \tag{1}$$

где $p(\mathbf{x}, y)$ — плотность совместного распределения. Выражение для $p(\mathbf{x}, y)$ можно получить, зная плотность условного распределения одной из величин и маргинальную плотность другой величины:

$$p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|y)p(y). \tag{2}$$

Эта процедура называется факторизацией. Результат факторизации зависит от способа разбиения плотности $p(\mathbf{x}, y)$.

Рассмотрим выборку $D = \{(\mathbf{x}_i, y_i)\}$, $i = 1, \dots, m$, состоящую из m реализаций величины $Z = (\mathbf{x}, y)$.

Предположим, что математическое ожидание величины $\mathbf{x}$ зависит от y, и распределение $\mathbf{x}$ есть смесь гауссовских распределений:

$$p(\mathbf{x}|y) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_y) - \frac{n}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma|\right), \text{ с вероятностью } p_y,\ y \in \{0, 1\}.$$

Пусть $p_1 = P$, $p_0 = 1 - P$. Тогда, воспользовавшись следующим свойством плотности совместного распределения:

$$\sum_y p(\mathbf{x}, y) = p(\mathbf{x}),$$

и вторым равенством в (2), получаем

$$p(\mathbf{x}) = Pp(\mathbf{x}|1) + (1 - P)p(\mathbf{x}|0).$$

Рассмотрим отношение

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{Pp(\mathbf{x}|1)}{(1 - P)p(\mathbf{x}|0)} = \frac{P}{1 - P} \exp\left(c - \mathbf{w}_1^T \mathbf{x} - \mathbf{x}^T \mathbf{w}_2\right), \tag{3}$$

где параметры c, $\mathbf{w}_0$ и $\mathbf{w}_1$ выражаются через параметры нормального распределения следующим образом:

$$c = \frac{1}{2}\left(\boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1\right),$$

$$\mathbf{w}_1 = \frac{1}{2}\left(\mu_{\mathbf{1}}{}^T \Sigma^{-1} - \mu_{\mathbf{0}}{}^T \Sigma^{-1}\right),$$

$$\mathbf{w}_2 = \frac{1}{2}\left(\Sigma^{-1} \mu_{\mathbf{1}} - \Sigma^{-1} \mu_{\mathbf{0}}\right).$$

Так матрица Σ симметрична, выражение (3) принимает вид:

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{P}{1-P} \exp\left(c - \mathbf{w}^T \mathbf{x}\right), \quad \text{где} \quad \mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2.$$

Учитывая, что $p(1|\mathbf{x}) + p(0|\mathbf{x}) = 1$, получаем

$$p(y|\mathbf{x}) = \left(1 + \left(\frac{1-P}{P}\right)^{2y-1} \exp\left(-(2y-1)\mathbf{w}^T \mathbf{x} + c\right)\right)^{-1}, \; y \in \{0, 1\}.$$

Тогда, воспользовавшись первым равенством в (2), получаем выражение для совместной плотности смешанного распределения:

$$p(\mathbf{x}, y) = p(y|\mathbf{x})(Pp(\mathbf{x}|1) + (1-P)(\mathbf{x}|0)). \tag{4}$$

Эта плотность выведена из определения условной плотности. Таким образом, если сделанные о характере условных распределений предположения верны, то значение $p(\mathbf{x}, y)$, вычисленное по (4) есть истинное значение плотности в точке $(\mathbf{x}, y)$. Поэтому в вычислительном эксперименте будем рассматривать оценку (4) как наиболее приближенную к истинному распределению.

Непараметрическая оценка плотности совместного распределения дискретной и непрерывной случайных величин

Может возникнуть ситуация, когда предположения о распределении случайных величин отсутствуют, либо по каким-либо причинам не могут быть использованы. Например, при классификации малых выборок использование плотности совместного распределения зависимых величин и независимых величин вместо плотности условного распределения зависимых переменных может улучшить качество классификации. Однако для этого нужно, чтобы плотность была вычислена без использования предположений об условных распределениях зависимых и независимых величин.

При непараметрическом оценивании плотности случайной величины Z в некоторой точке z, производится усреднение частоты появления в выборке ближайших к ней точек выборки z_i. При этом используются ядерные функции, убывающие с увеличением расстояния между z и Z_i. Обозначим ядерную функцию $K(u)$, тогда оценка плотности величины Z в точке z может быть получена с помощью:

$$\hat{p}(z) = \frac{1}{m} \sum_{i=1}^{m} K(z - z_i). \tag{5}$$

Определим ядерную функцию смешаной случайной величины $Z = (\mathbf{x}, y)$ в точке $z = (\mathbf{s}, t)$ как произведение дискретного и непрерывного ядер:

$$K_{h,\lambda}(z - z_i) = L_\lambda(t - y_i)C_h\left(\frac{\mathbf{s} - \mathbf{x}_i}{h}\right).$$

Ядерная функция для дискретной переменной имеет вид

$$L_\lambda(u) = \begin{cases} \lambda, u = 0, \\ 1 - \lambda, u \neq 0. \end{cases}$$

Ядерная функция $C_h(\mathbf{u})$ для вектора непрерывных переменных также определяется как произведение одномерных ядер. Пусть смешанная случайная величина Z включает в себя n непрерывных случайных величин, тогда

$$C_h(\mathbf{u}) = \frac{1}{h^m}\prod_{j=1}^{n} c(u_j).$$

В качестве $c(u)$ выберем ядро Епанечникова:

$$c(u) = \frac{3}{4}(1 - u^2)[|u| < 1],$$

где $[|u| < 1]$ — индикаторная функция. Ядро Епанечникова не учитывает влияние точек, отстоящих от $\mathbf{s}$ дальше, чем на h, и удовлетворяет условию

$$\int_{-\infty}^{\infty} c(u)\mathrm{d}u = 1.$$

Кроме того, оно минимизирует среднеквадратичную ошибку аппроксимации.

Выбор параметров сглаживания

Рассмотрим интегральное среднеквадратичное отклонение (MISE) полученной оценки $\hat{p}(z)$ от истинной плотности распределения $p(z)$. Подразумевая под $\int dz$ суммирование по дискретным переменным и интегрирование по непрерывным, запишем выражение для MISE

$$E_{\text{MISE}} = \int (\hat{p}(z) - p(z))^2 dx = \int \hat{p}^2(z)dx - 2\int \hat{p}(z)p(z)dz + \int p^2(z)dz. \tag{6}$$

Второй член суммы в правой части (6) есть математическое ожидание величины $\hat{p}(Z)$, его можно оценить как

$$\int \hat{p}(z)p(z)dz \approx \frac{1}{m}\sum_{i=1}^{m}\hat{p}_{\text{LOO}}(z_i) = \frac{1}{m(m-1)}\sum_{i_1=1}^{m}\sum_{i_2=1,\ i_2\neq i_1}^{m} K_{h,\lambda}(z_{i_1} - z_{i_2}).$$

Здесь использована оценка скользящего контроля

$$\hat{p}_{\text{LOO}}(z_{i_1}) = \frac{1}{m-1}\sum_{i_2=1,\ i_2\neq i_1}^{m} K_{h,\lambda}(z_{i_1} - z_{i_2}).$$

Последний член в (6) можно опустить, так как он не зависит от выбора ширины окна h и параметра сглаживания λ. Получаем:

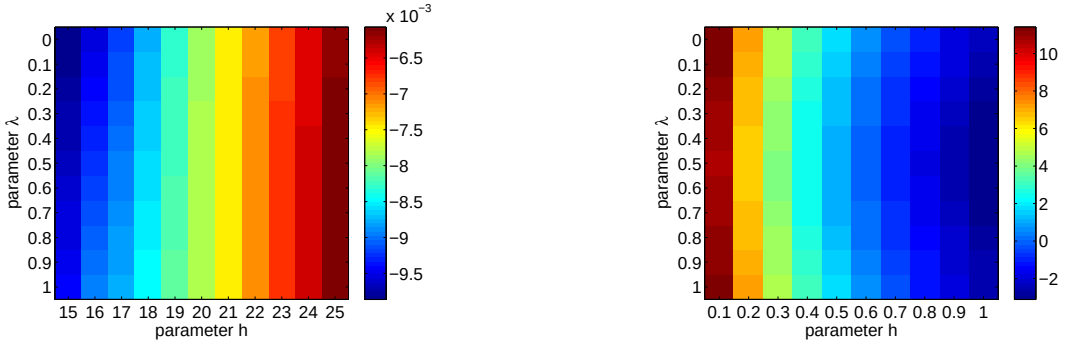
$$\tilde{E}_{\text{MISE}} = \frac{1}{m^2}\sum_{i_1=1}^{m}\sum_{i_2=1}^{m} K^{(2)}_{h,\lambda}(z_{i_1}, z_{i_2}) - \frac{2}{m(m-1)}\sum_{i_1=1}^{m}\sum_{i_2=1,\ i_2\neq i_1}^{m} K_{h,\lambda}(z_{i_1} - z_{i_2}), \tag{7}$$

где $K^{(2)}_{h,\lambda}(z_{i_1}, z_{i_2}) = L^{(2)}_{\lambda}(y_{i_1}, y_{i_2})C^{(2)}_{h}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$, а выражения для $L^{(2)}_{\lambda}(y_{i_1}, y_{i_2})$ и $C^{(2)}_{\lambda}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$ определяются как

$$L^{(2)}_{\lambda}(y_{i_1}, y_{i_2}) = \sum_{y\in D} L_{\lambda}(y - y_{i_1})L_{\lambda}(y - y_{i_2}),$$

$$C^{(2)}_{\lambda}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sum_{\mathbf{x}\in D} C_{\lambda}(\mathbf{x} - \mathbf{x}_{i_1})C_{\lambda}(\mathbf{x} - \mathbf{x}_{i_2}).$$

Метод оценки параметров λ и h, основанный на минимизации выражения (6), называется кросспроверкой.



a. b.

Рис. 1: Значения $\tilde{E}_{\text{MISE}}$ в зависимости от λ и h при больших значениях h (a) и при различных значениях h в логарифмическом масштабе (b).

Об оптимальности ядра Епанечникова

Точность оценки функции $p(z)$ зависит не только от выбора ширины окна h и параметра сглаживания λ, но и от выбора ядерной функции $K(u)$. Рассмотрим поведение среднеквадратичной ошибки оценки $\hat{p}(z)$ как функции ядра K. В работе [10] показано, что для минимизации среднеквадратичного отклонения по ядру K необходимо минимизировать произведение

$$I_{\text{MSE}}(K) = \left(\int K(u)^2 du\right)^2 \int u^2 K(u) du.$$

При этом ядерная функция должна удовлетворять следующим требованиям:

$$\int K(u)du = 1, \tag{8}$$

$$K(u) = K(-u), \tag{9}$$

$$\int u^2 K(u)du = 1. \tag{10}$$

Ограничение (10) снижает вес элементов с $u \approx 0$, препятствуя выбору чересчур узких функций. Учитывая (10), приходим к задаче минимизации $\int K(u)^2 du$ при ограничениях (8), (9), (10). Запишем лагранжиан этой задачи:

$$L(K, \mu_1, \mu_2) = \int K(u)^2 du + \mu_1 \left(\int K(u) du - 1 \right) + \mu_2 \left(\int u^2 K(u) du - 1 \right).$$

В точке экстремума K_0 вариация лагранжиана должна равняться нулю. Обозначив через $\Delta K = K - K_0$ малое отклонение от экстремальной функции, получим:

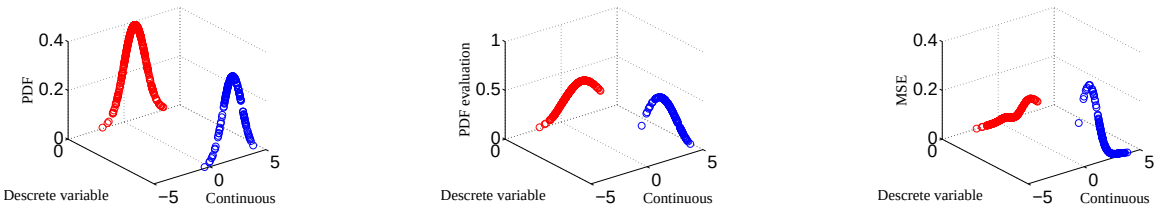
$$2 \int \Delta K(u)(2K(u) + \mu_1 + \mu_2 u^2) du = 0,$$

и, следовательно,

$$2K(u) + \mu_1 + \mu_2 u^2 = 0.$$

Заметим, что $K(u)$ обращается в ноль при $u = \pm \left(-\frac{\mu_1}{\mu_2} \right)^2$. Выбрав ядро с носителем $|u| < \left(-\frac{\mu_1}{\mu_2} \right)^2$, и учитывая условия (8), (9), (10), получаем ядро перенормированное Епанечникова

$$K(u) = \frac{3}{4 \cdot \sqrt{15}} \left(1 - \frac{u^2}{15} \right) \left[\frac{|u|}{\sqrt{15}} < 1 \right].$$



a. b. c.

Рис. 2: a. Распределение тестовых данных $p(\mathbf{x}, y)$, восстановленное факторизацией. Так как условные распределения известны, полученная оценка совпадает с истинным совместным распределением. b. Непараметрическая оценка функции $\hat{p}(\mathbf{x}, y)$. c. Зависимость $E_{\text{MSE}}(\mathbf{x}, y)$, среднеквадратичного отклонения $\hat{p}$ от p.

Вычислительный эксперимент

В вычислительном эксперименте рассмотрена выборка синтетических данных $D = \{z_i\} = \{(\mathbf{x}_i, y_i)\}$, $i = 1, \ldots, m$, порожденная таким образом, что:

$$p(y = 1) = P, \quad p(y = 0) = 1 - P, \tag{11}$$

$$p(\mathbf{x}|y) = \mathcal{N}(\boldsymbol{\mu}_y, \sigma_y^2 I). \tag{12}$$

Этот вид смешанной случайной величины описан в разделе 4. Заметим, что если предположения (11), (12) выполняются, то плотность распределения, полученная с помощью (4), есть истинная плотность. Таким образом, для синтетических данных истинная плотность совместного распределения известна. В случае тестовых данных, предположения могут быть слишком грубы, и восстановленная с помощью факторизиции плотность совместного распределения будет отличаться от истинной плотности.

Будем рассматривать функции плотности $\tilde{p}(z)$ и $\hat{p}(z)$, восстановленные с помощью (4) и (5). Графики этих функций изображены на рисунке 4. Для оценки точности восстановления плотности будем использовать ошибку $\tilde{E}_{\text{MSE}}$:

$$\tilde{E}_{\text{MSE}}(z_i) = \left(\hat{p}(z_i) - \tilde{p}(z_i)\right)^2. \tag{13}$$

Оценки функции $p(z)$, полученные различными способами, будем сравнивать, используя

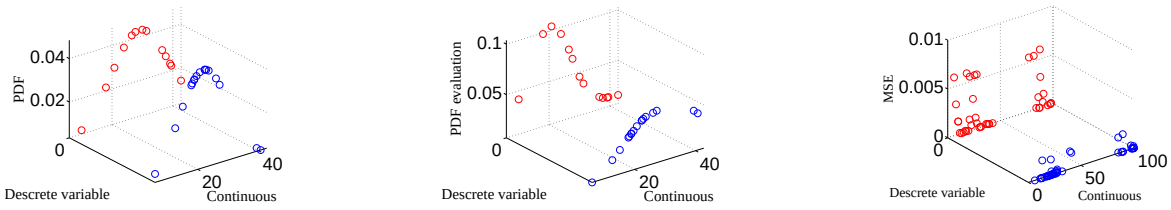
a. b.

Рис. 3: a. Зависимость расстояния Кульбака-Лейблера $D_{KL}(p|\hat{p})$ от параметра сглаживания h. b. Зависимость расстояния Кульбака-Лейблера $D_{KL}(p|\hat{p})$ от параметра λ при оптимальном значении h.

расстояние Кульбака-Лейблера:

$$D_{KL}(p, \hat{p}) = \int_z p(z) \log \frac{p(z)}{\hat{p}(z)} dz = \mathbb{E}\left[\frac{p(Z)}{\hat{p}(Z)}\right]. \tag{14}$$

Рассмотрим зависимости расстояния между восстановленными плотностями $D_{KL}(\tilde{p}, \hat{p})$ от параметров сглаживания λ, h. Эти зависимости и результаты кросспроверки изображены на рисунке 4. Оба способа приводят к выбору одного значения параметра λ, но дают различные оценки параметра h. Будем использовать расстояние Кульбака-Лейблера, так как его проще вычислять.

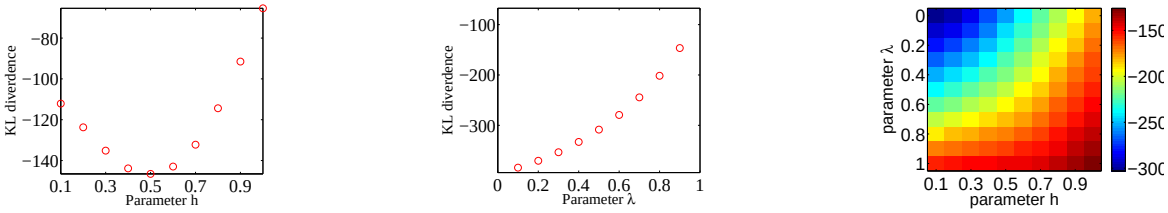


a. b. c.

Рис. 4: a. Оценка совместного распределения реальных данных $\tilde{p}(\mathbf{x}, y)$, полученная с помощью факторизации. В отличии от случая с известными условными распределениями, эта оценка может не совпадать с истинной плотностью совместного распределения. Ее точность зависит от точности вероятностных предположений. b. Непараметрическая оценка функции $\hat{p}(\mathbf{x}, y)$. c. Зависимость $E_{\text{MSE}}(\mathbf{x}, y)$, среднеквадратичного отклонения $\hat{p}$ от p.

Рассмотрим зависимости расстояния между восстановленными плотностями $D_{KL}(\tilde{p}, \hat{p})$ от параметров сглаживания λ, h. Эти зависимости и результаты кросспроверки изображены на рисунке 4. Оба способа приводят к выбору одного значения параметра λ, но дают

различные оценки параметра h. Заметим, расстояние Кульбака-Лейблера проще вычислять, однако его использование возможно только при известном истинном распределении.
.



a. b. c.

Рис. 5: a. Зависимость расстояния Кульбака-Лейблера $D_{KL}(\tilde{p}|\hat{p})$ от параметра сглаживания h. b. Зависимость расстояния Кульбака-Лейблера $D_{KL}(\tilde{p}|\hat{p})$ от параметра λ при оптимальном значении h. c. Зависимость $\tilde{E}_{\text{MISE}}$ от параметров h и λ.

Рассмотрим также выборку реальных данных. Будем оценивать $p(z)$, предполагая (11) и (12). Функции $\tilde{p}(z)$, $\hat{p}(z)$ и оценка E_{MSE} изображены на рисунке 4. Видно, что предположения оказались слишком сильны, условные распределения $p(\mathbf{x}|1)$ и $p(\mathbf{x}|0)$ не являются нормальными, и функция $\tilde{p}(z)$ далека от истинной функции распределения.

Заключение

В работе рассматривается задача оценки плотности совместного неоднородного распределения. Выборка состоит их дискретных и непрерывных величин случайных, и имеет малый объем, в следствие чего не удается сделать предположений об условных или маргинальных распределениях этих величин. Предложен способ оценки плотности совместного распределения, основанный на применении методов ядерного сглаживании. В частности, для непрерывных величин используется ядро Епанечникова, а для дискретных — сглаженная индикаторная функция. Для сравнения рассмотрен метод факторизации совместного распределения, то есть разбиение его на условное и маргинальное распределение подмножества случайных величин, образующих многомерную случайную величину, плотность которой необходимо оценить.

Литература

[1] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001.

[2] Rennie J., Shih L., Teevan J., and Karger D. Tackling The Poor Assumptions of Naive Bayes Classifiers. Proceedings of the Twentieth International Conference on Machine Learning (ICML). 2003.

[3] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77-2(1989), 257–285.

[4] Stamp, M. A Revealing Introduction to Hidden Markov Models. San Jose State University, 2012.

[5] Everitt, B. S. The Cambridge Dictionary of Statistics . Cambridge University Press, 2002.

[6] Olkin, I. and Tate, R. F. Annnals of Math. Statistics, 36-1(1965), 343-344.

[7] McCulloch, C. E. Joint modeling of mixed outcome types using latent variables. Statistical Methods in Medical Research, 17(2008), 53–73.

[8] Nelsen R. B. An introduction to copulas. Springer, 2006.

[9] Charpentier A., Fermanian J.-D., Scaillet O. The Estimation of Copulas: Theory and Practice, 2006.

[10] Хардле, В. Прикладная непараметрическая регрессия. Москва "Мир 1993.

[11] Li, Q., Racine, J. Nonparametric Estimation of Distributions with Categorical and Continuous Data, Journal of Multivariate Analysis, 86-2(2003), 266 - 292.

Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании*

В. Р. Целых[1]***, К. В. Воронцов***[2]
celyh@inbox.ru, voron@forecsys.ru
1 — Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»
2 — Вычислительный центр РАН

Критерий согласия Пирсона неприменим к сильно разреженным распределениям, так как в этих случаях распределение статистики плохо описывается асимптотическим законом хи-квадрат, зависит от объема выборки и вида исходного распределения. В данной работе предлагаются статистические критерии, основанные на вычислении эмпирических распределений статистик путем сэмплирования. Рассматривается их применение в задачах анализа текстов, в частности, для проверки гипотезы условной независимости при построении и оценивании вероятностных тематических моделей.

Ключевые слова: *критерий согласия, статистика хи-квадрат, сэмплирование, метод Монте-Карло, закон Ципфа, вероятностная тематическая модель, гипотеза условной независимости.*

Goodness-of-fit tests for sparse multinomial distributions with application to topic modeling*

V. R. Tselykh[1]***, K. V. Vorontsov***[2]
1 — Moscow Institute of Physics and Technology
2 — Computing Center of the Russian Academy of Sciences

Pearson's goodness-of-fit test is not appropriate for sparse multinomial distributions. In this case, the distribution of statistic is not asymptotically chi-squared, it depends on a sample size and on a form of the tested distribution. The article suggests statistical criteria based on empirical distribution of a statistic obtained from sampling. Their application to text analysis is considered, in particular, to testing the conditional independence hypothesis for probabilistic topic models evaluation.

Keywords: *goodness-of-fit test, chi-squared statistics, sampling, Zipf's law, probabilistic topic model, conditional independence.*

Введение

Стандартные критерии согласия для дискретных распределений плохо подходят, когда число возможных значений наблюдаемой переменной значительно превосходит число наблюдений либо когда многие значения имеют близкие к нулю вероятности [1, 2]. Такие распределения называют разреженными. В этих случаях распределение статистики не описывается классической асимптотикой, может зависеть от объема выборки и степени разреженности исходного распределения.

Разреженные распределения возникают в задачах статистического анализа текстов, когда текст рассматривается как дискретное распределение на множестве слов, и требу-

Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.

ется проверить, является ли заданный фрагмент текста случайной выборкой из известной генеральной совокупности. В данной работе рассматривается ситуация, когда генеральная совокупность фиксирована, и требуется проверять много фрагментов. Тогда становятся оправданными методы сэмплирования, выполняющие большой объем вычислений на этапе предварительной обработки генеральной совокупности.

Предлагаются непараметрические статистические тесты, основанные на сэмплировании с возвращением и без возвращения, а также параметрический метод, предназначенный для проверки согласия с законом Ципфа. Для параметрического метода строится регрессионная модель, выражающая квантиль распределения через параметры задачи. Преимущество регрессионного теста в том, что, в отличие от методов сэмплирования, его не нужно перестраивать заново для каждого распределения.

Рассматривается применение предложенных статистических тестов для проверки гипотезы условной независимости при построении и оценивании вероятностных тематических моделей коллекций текстовых документов.

Критерий согласия хи-квадрат

Пусть имеется выборка n независимых наблюдений $\{x_1, \ldots, x_n\}$ случайной величины, принимающей значения из конечного множества Ω. Ее эмпирическое распределение определяется как доля наблюдений x_i, равных x:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}[x_i = x], \quad x \in \Omega.$$

Критерий хи-квадрат проверяет гипотезу о том, что случайная величина имеет заданное распределение $p(x)$, $x \in \Omega$. Для этого вычисляется статистика хи-квадрат:

$$X^2 = n\sum_{x\in\Omega}\frac{(\hat{p}(x) - p(x))^2}{p(x)}. \tag{1}$$

Распределение статистики X^2 стремится к распределению хи-квадрат с $k = |\Omega| - 1$ степенями свободы: $X^2 \sim \chi^2(k)$. Нулевая гипотеза отвергается на уровне значимости α, если значение статистики превышает $(1-\alpha)$-квантиль этого распределения: $X^2 > \chi^2_{1-\alpha}(k)$.

Считается, что асимптотика хи-квадрат применима, если объем выборки $n \geqslant 50$ и ожидаемое число наблюдений $np(x) \geqslant 5$ для каждого $x \in \Omega$. В случаях *разреженных* распределений $p(x)$, когда вероятности $p(x)$ малы для многих $x \in \Omega$ или когда $|\Omega| \gg n$, второе условие может не выполняться даже на очень больших выборках [1]. Стандартная рекомендация — объединять значения $x \in \Omega$ в группы — для сильно разреженных распределений оказывается неприемлемой, так как результат существенно зависит от способа группирования, который выбирается произвольно.

В качестве иллюстрации рассмотрим распределение, называемое *законом Ципфа*:

$$p(x) = Ax^{-s}, \quad x \in \Omega = \{1, \ldots, v\}, \tag{2}$$

где $A = \left(\sum_{x=1}^{v} x^{-s}\right)^{-1}$ — нормировочный множитель, s — параметр. Этот закон неплохо описывает частоты слов в текстах на естественных языках, если за x принимать номера слов, упорядоченных по убыванию частоты. Параметр s зависит от языка и от корпуса текстов, по которому делается оценка, но в большинстве экспериментов значение s близко к 1 и находится в пределах от 0.9 до 1.2 [3, 4].

Чем больше значение параметра s и размер словаря v, тем более разрежено распределение $p(x)$. Проведем простой вычислительный эксперимент. Возьмем типичные значения параметра $s = 1$ и размера словаря $v \in \{50,\ 500,\ 1000,\ 5000\}$. Сгенерируем $N = 1000$ выборок (искусственных текстов) длины $n = 100$ из распределения (2), и для каждой выборки вычислим значение статистики X^2.

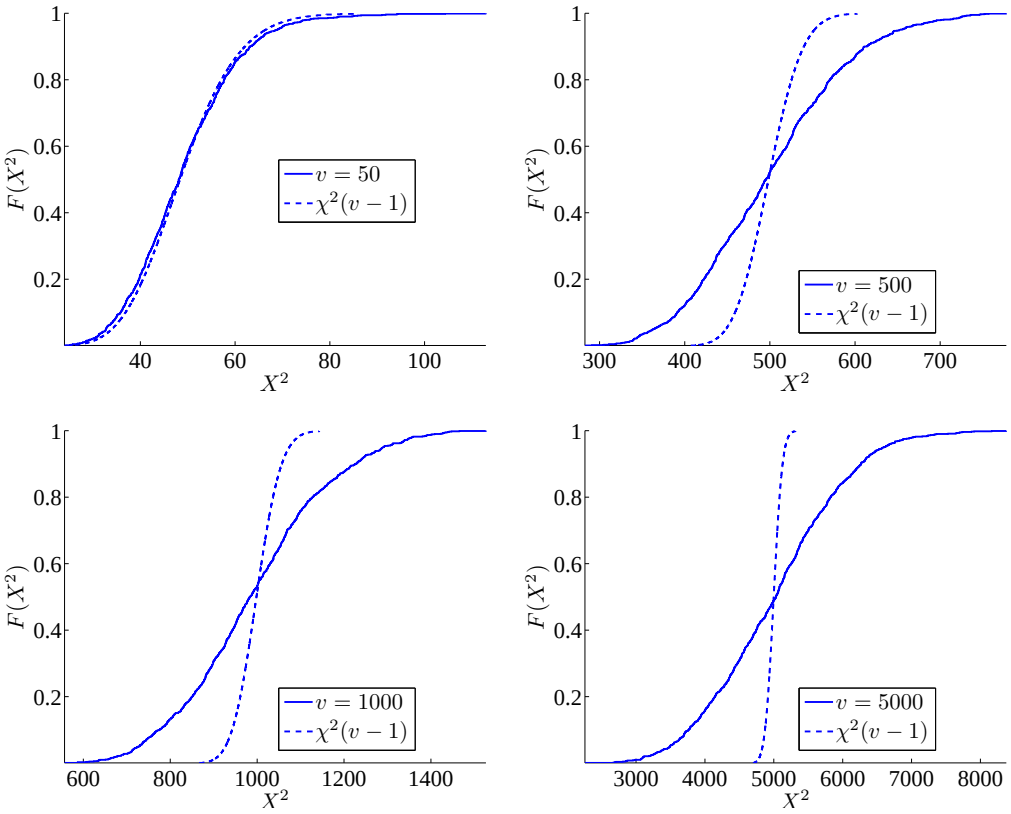


Рис. 1: Функции распределения статистики X^2 при $s = 1$, $v = 50,\ 500,\ 1000,\ 5000$, $n = 100$, $N = 1000$ и соответствующие функции распределения $\chi^2(v-1)$.

На рис. 1 сплошными линиями показаны эмпирические распределения статистики X^2, пунктирными линиями — распределения $\chi^2(v-1)$. Чем больше размер словаря, тем сильнее разрежено распределение $p(x)$ и тем сильнее отличаются $(1-\alpha)$-квантили этих распределений (при типичном значение $\alpha = 0.05$).

Таким образом, распределение хи-квадрат не может быть использовано в практических задачах анализа текстов, когда требуется проверить, является ли заданный текст $\hat{p}(x)$ случайной выборкой из корпуса текстов $p(x)$.

Тест на основе сэмплирования

Для разреженных распределений $p(x)$ предлагается вместо асимптотического распределения $\chi^2(k)$ статистики X^2 использовать эмпирическое распределение.

Построение теста. Генерируется N независимых выборок объема n из заданного дискретного распределения $p(x)$. Для каждой выборки вычисляется эмпирическое распределение $\hat{p}_j(x)$, $j = 1, \ldots, N$ и значение статистики X_j^2 по формуле (1). По полученным значениям $X_1^2, \ldots, X_N^2$ строится эмпирическая функция распределения статистики

$$\hat{F}_n(X^2) = \frac{1}{N} \sum_{j=1}^{N} [X^2 > X_j^2]$$

и вычисляется ее $(1-\alpha)$-квантиль $\hat{F}_{n,1-\alpha}$. Число N рекомендуется брать не менее 1000, если необходимо оценивать всю функцию распределения. Однако если оценивается только одна квантиль, N можно брать порядка нескольких десятков [2].

Применение теста. Пусть задана выборка объема n, по которой построено эмпирическое распределение $\hat{p}(x)$ и вычислено значение статистики X^2 согласно (1). Если

Алгоритм 1 Построение теста путем рекуррентного вычисления значений статистики X^2 по N одновременно растущим выборкам объема n.

Вход: $p(x)$, N, $n_{\max}$, α;
Выход: $\hat{F}_{n,1-\alpha}$ для всех $n = 1, \ldots, n_{\max}$;

для всех $j := 1, \ldots, N$
  сэмплировать первый элемент j-й выборки $\xi \sim p(x)$;
  инициализировать эмпирическую гистограмму для j-й выборки: $H_j(x) := [x = \xi]$ для всех $x \in \Omega$;
  инициализировать значение статистики X^2 для j-й выборки: $X^2_{j1} := 1/p(\xi) - 1$;
для всех $n := 1, \ldots, n_{\max} - 1$
  для всех $j := 1, \ldots, N$
    сэмплировать $(n+1)$-й элемент j-й выборки $\xi \sim p(x)$;
    обновить эмпирическую гистограмму для j-й выборки: $H_j(\xi) := H_j(\xi) + 1$;
    обновить значение статистики X^2 для j-й выборки:
$$X^2_{j,n+1} := \frac{nX^2_{j,n} + 1}{n+1} + \frac{2H_j(\xi) - 1}{(n+1)p(\xi)} - 2;$$
для всех $n := 1, \ldots, n_{\max}$
  упорядочить $X^2_{1,n}, \ldots, X^2_{N,n}$ по возрастанию;
  $\hat{F}_{n,1-\alpha} := X^2_{N(1-\alpha),n}$;

$X^2 > \hat{F}_{n,1-\alpha}$, то нулевая гипотеза о том, что данная выборка порождена распределением $p(x)$, отклоняется.

Рекуррентное построение теста. Как будет показано ниже, в случае разреженных распределений значение квантили $\hat{F}_{n,1-\alpha}$ может зависеть от объема выборки n. Строить тест заново для каждой выборки довольно накладно. Поэтому предлагается рекуррентный метод, позволяющий при заданном распределении $p(x)$ вычислить квантили для всех значений n один раз, и затем быстро осуществлять проверку нулевой гипотезы для выборок различного объема n.

В рекуррентном методе N выборок $\{x_{j1}, \ldots, x_{jn}\}$ наращиваются одновременно, где $j = 1, \ldots, N$ — номер выборки, $n = 1, \ldots, n_{\max}$ — объем выборки. Для каждого j строится эмпирическая гистограмма $H_j(x) = n\hat{p}_j(x)$. При добавлении каждого нового наблюдения $\xi = x_{j,n+1}$, сэмплированного из распределения $p(x)$, обновляется гистограмма и пересчитывается значение статистики $X^2_{j,n+1}$ по значению $X^2_{j,n}$. Сэмплированные выборки не сохраняются. В процессе работы алгоритм формирует двумерный массив значений статистики $X^2_{j,n}$ и одномерный массив эмпирических гистограмм $H_j(x)$. В случае $|\Omega| \gg n_{\max}$ для хранения эмпирических гистограмм лучше использовать специальные структуры данных — разреженные векторы, не выделяющие память под нулевые значения $H_j(x)$. В таком случае расход памяти для данного алгоритма составляет $O(n_{\max}N)$; вычислительная сложность $O(n_{\max}N \log N)$. Детали реализации показаны в Алгоритме 1.

Регрессионный тест

Рассмотрим частную постановку задачи: проверяется нулевая гипотеза о том, что выборка с эмпирическим распределением $\hat{p}(x)$ порождена распределением Ципфа (2) с параметром s. Будем строить распределение статистики X^2 с помощью сэмплирования и исследовать зависимость квантиля $\hat{F}_{n,1-\alpha}$ от параметров n, s и v.

На рис. 2 показана зависимость 0.95-квантиля от объема выборки n и ее интерполяция функцией $\tilde{F}_{1-\alpha}(n) = A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4}$ с параметрами A, B, C, D, E.

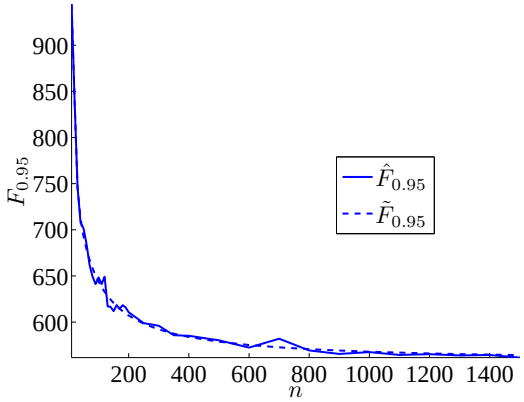


Рис. 2: Зависимость 0.95-квантиля X^2 от объема выборки n при $s = 1$, $v = 500$, $N = 1000$ и ее интерполяция.

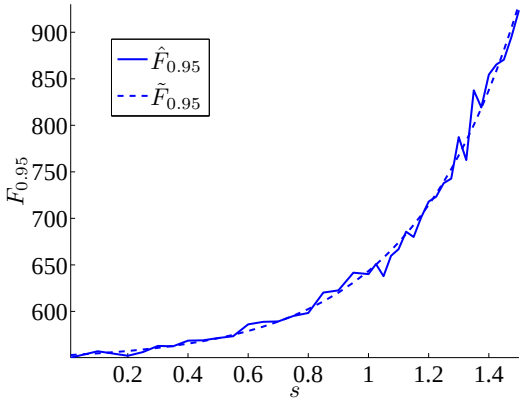


Рис. 3: Зависимость 0.95-квантиля X^2 от параметра s при $n = 100$, $v = 500$, $N = 1000$ и ее интерполяция.

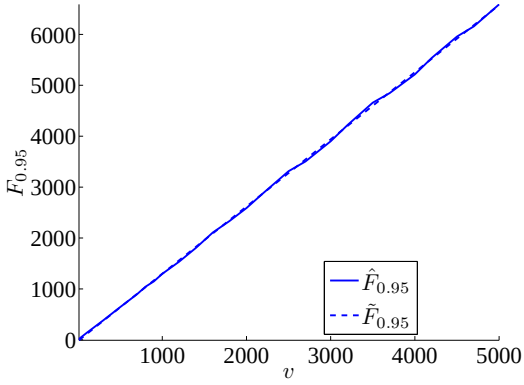


Рис. 4: Зависимость 0.95-квантиля X^2 от $v = |\Omega|$ при $s = 1$, $n = 100$, $N = 1000$ и ее интерполяция.

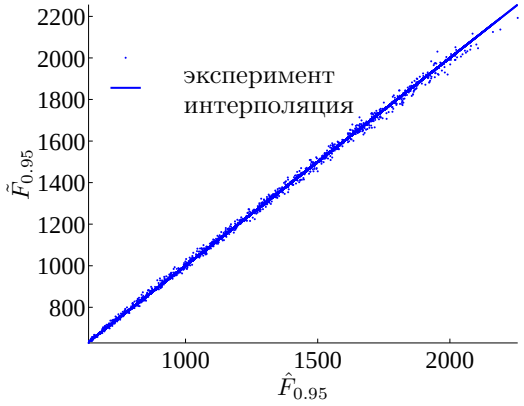


Рис. 5: Зависимость 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{1-\alpha}$, от их эмпирических значений при различных s, n, v.

На рис. 3 показана зависимость 0.95-квантиля от показателя s в законе Ципфа и ее интерполяция функцией $\tilde{F}_{1-\alpha}(s) = F + GH^s$ с параметрами F, G, H.

На рис. 4 показана зависимости 0.95-квантиля от параметра $v = |\Omega|$ и ее линейная интерполяция $\tilde{F}_{1-\alpha}(v) = I + Jv$ с параметрами I, J.

Построение регрессионного теста. Чтобы найти общий вид зависимости $\tilde{F}_{1-\alpha}(s, v, n)$, применим эмпирический подход. Сформируем обучающую выборку из 1000 троек (s, v, n), равномерно выбранных из параллелепипеда $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки вычислим значение $\hat{F}_{n,0.95}$.

Для поиска нелинейной регрессионной зависимости используем алгоритм символьной регрессии MVR-composer [5, 6]. Преимущество этого алгоритма в том, что он автоматически подбирает формулу регрессии среди всевозможных суперпозиций заданного множества элементарных функций. В нашем случае MVR-composer находит следующую модель регрессии: $\tilde{F}^1_{1-\alpha}(s, v, n) = (A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(F + GH^s)(I + Jv)$ и определяет оптимальные значения 10 параметров $A, B, C, D, E, F, G, H, I, J$. Рассмотрим также

некоторые упрощения этой модели:

$$\tilde{F}^2_{1-\alpha}(s,v,n) = A(1 + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^3_{1-\alpha}(s,v,n) = A(1 + Bn^{-c})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^4_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c})(1 + GH^s);$$
$$\tilde{F}^5_{1-\alpha}(s,v,n) = Av(1 + GH^s);$$
$$\tilde{F}^6_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c}).$$

Параметры этих моделей настроим с помощью функции nlinfit программы Matlab. Начальные приближения всех параметров положим равными 1, кроме параметра A, который инициализируем средним значением $\hat{F}_{n,1-\alpha}/v$ по всей выборке. Получим следующие значения среднеквадратичной ошибки (СКО) на обучающей и контрольной выборках из 1000 случайных троек (s,v,n) каждая:

модель	$\tilde{F}^1_{1-\alpha}$	$\tilde{F}^2_{1-\alpha}$	$\tilde{F}^3_{1-\alpha}$	$\tilde{F}^4_{1-\alpha}$	$\tilde{F}^5_{1-\alpha}$	$\tilde{F}^6_{1-\alpha}$
число параметров	10	8	6	5	3	3
СКО (обучение)	16.3	16.8	16.8	16.7	52.2	43.7
СКО (контроль)	15.8	16.1	16.0	16.0	50.9	43.8

Сравнение СКО на обучающей и контрольной выборках показывает, что переобучения нет ни в одной из моделей. Модель $\tilde{F}^4_{1-\alpha}$ представляется оптимальной по точности и числу параметров. Дальнейшее упрощение модели приводит к резкому увеличению СКО. Оптимальные значения параметров для нее: $A = 0.913$, $B = 3.98$, $c = 0.636$, $G = 0.00458$, $H = 36.8$.

На рис. 4 показан график зависимости 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{0.95}$, от их эмпирических значений при различных s, n, v. Сплошной линией изображена «идеальная» прямая $\tilde{F} = \hat{F}$.

Таким образом, в отличие от классического критерия хи-квадрат квантиль распределения статистики X^2 существенно зависит от объема выборки n и от вида распределения $p(x)$, в частности, от показателя степени s в законе Ципфа, отвечающего за разреженность распределения. Построенная регрессионная модель довольно точно описывает зависимость 0.95-квантили от параметров s, n, v. Эту зависимость можно построить один раз, вместо того чтобы строить тест для каждого распределения $p(x)$. Предварительно необходимо убедиться, что распределение $p(x)$ описывается законом Ципфа, и найти значение параметра s. Данное обстоятельство сужает область применимости регрессионного теста.

Анализ качества регрессионного теста. Оценим вероятности ошибок первого и второго рода предложенного регрессионного теста в эксперименте.

Ошибкой первого рода называется отклонение нулевой гипотезы при условии ее истинности. Вероятность ошибки первого рода равна уровню значимости $\alpha = 0.05$. Для эксперимента сгенерируем контрольную выборку из 500 различных троек (s,v,n), равномерно распределенных на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки сгенерируем 1000 выборок объема n из распределения Ципфа $p(x)$ с параметрами v и s и вычислим значение статистики X^2. Оценим вероятность ошибки первого рода как долю выборок, для которых нулевая гипотеза отклонялась: $X^2 > \tilde{F}^4_{0.95}(s,v,n)$. Оценка вероятности ошибки первого рода составляет 0.0496 ± 0.0141 с доверительной вероятностью 0.95.

Ошибкой второго рода называется принятие гипотезы $H_0\colon p(x)$ при условии истинности ее альтернативы $H_1\colon p'(x)$. Вероятность ошибки второго рода существенно зависит от альтернативы — чем более похожи распределения $p(x)$ и $p'(x)$, тем больше вероятность ошибки. Исследуем способность теста различать распределения, отличающиеся на

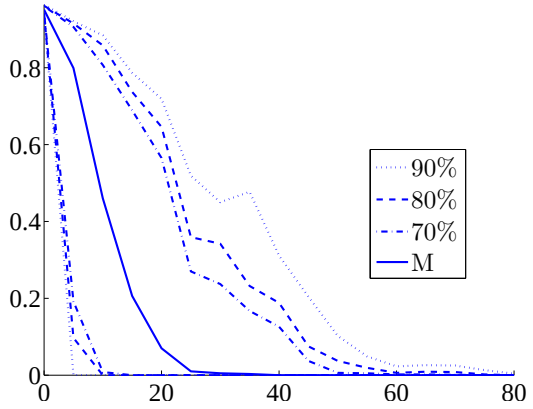


Рис. 6: Зависимость вероятности ошибки второго рода от K при $\mu = 0.01$.

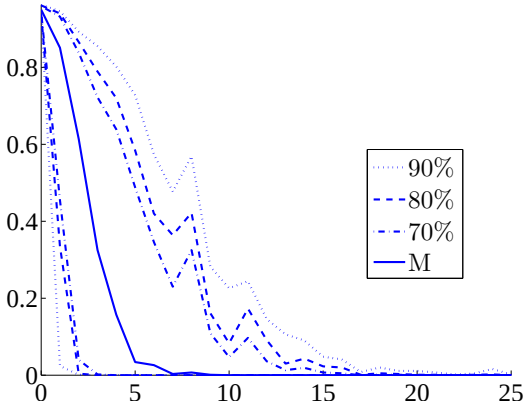


Рис. 7: Зависимость вероятности ошибки второго рода от K при $\mu = 0.05$.

небольшом числе элементов x из Ω. Выделим из множества $\Omega = \{1, \ldots, v\}$ подмножество элементов с наибольшими вероятностями: $\Omega_0 = \{x\colon p(x) > \mu p(1)\}$ при заданном $\mu \in (0, 1)$. Построим распределение $p'(x)$ из $p(x)$ следующим образом: выберем K различных случайных элементов множества Ω_0 и их вероятности поменяем местами с вероятностями K различных случайных элементов множества $\Omega \backslash \Omega_0$.

Из полученного распределения $p'(x)$ сгенерируем выборки, для каждой построим эмпирическое распределение $\hat{p}(x)$ и вычислим статистику X^2. Если $X^2 \leqslant \tilde{F}^4_{0.95}(s, v, n)$, то для данной выборки гипотеза H_0 ошибочно принимается. Долю выборок, при которых это происходит, примем в качестве оценки вероятности ошибки второго рода.

Для каждого K сгенерируем 200 различных троек (s, v, n) из равномерного распределения на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$ и вычислим 200 оценок вероятности ошибки второго рода. На рис. 6 и рис. 7 показаны зависимости медианы M и доверительных границ 90%, 80%, 70% вероятности ошибки второго рода от числа перестановок K при $\mu = 0.01$ и $\mu = 0.05$. По мере увеличения K распределения $p(x)$ и $p'(x)$ все сильнее отличаются, и вероятность ошибки второго рода уменьшается. По мере увеличения μ различия становятся менее контрастными, и вероятность ошибки второго рода убывает медленнее. При $\mu = 0.01$ она становится меньше 0.1 при $K = 20$, при $\mu = 0.05$ она достигает этого значения при $K = 5$.

Отсюда, в частности, можно сделать вывод, что различные тексты, отличающиеся лишь 5 высокочастотными терминами, в среднем довольно надежно различаются по их случайным фрагментам.

Вероятностные тематические модели

Тематическое моделирование (topic modeling) — одно из активно развивающихся приложений машинного обучения к анализу текстов [7]. *Тематическая модель* коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие термины образуют каждую тему. *Вероятностная тематическая модель* описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Это позволяет решать задачи классификации, кластеризации и категоризации текстов, а также создавать тематические поисковые системы, позволяющие по тексту произвольной длины находить документы схожей тематики.

Исходными данными для тематической модели является множество (коллекция) текстовых документов D и множество (словарь) терминов W. Каждый документ $d \in D$ представляется последовательностью терминов $(w_1, \ldots, w_{n_d})$ из W, где n_d — длина документа. Через n_{dw} обозначается число вхождений термина w в документ d.

Вероятностные модели основаны на следующих предположениях [8, 9].

Во-первых, предполагается, что для выявления тематики достаточно знать, какие термины встречаются в каких документах, но не важен ни порядок терминов в документах (*гипотеза «мешка слов»*), ни порядок документов в коллекции (*гипотеза «мешка документов»*). Другими словами, предполагается, что тематику документа можно узнать даже после случайной перестановки терминов, хотя для человека такой текст теряет смысл.

Во-вторых, предполагается, что существует конечное множество тем T и дискретное распределение $p(d, w, t)$ на $D \times W \times T$, порождающее последовательность независимых наблюдений — троек (d_i, w_i, t_i), $i = 1, \ldots, n$. Переменная t является латентной (скрытой), и наблюдаемая коллекция документов представляет собой последовательность пар (d_i, w_i), $i = 1, \ldots, n$, оставшихся после отбрасывания всех тем.

В-третьих, предполагается, что условное распределение вероятностей терминов $p(w \mid d, t)$ в любом документе d зависит только от темы t, но не от самого документа. Это предположение называется *гипотезой условной независимости*:

$$p(w \mid d, t) = p(w \mid t). \tag{3}$$

Согласно формуле полной вероятности и гипотезе условной независимости,

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d). \tag{4}$$

Построить тематическую модель коллекции — означает по известной левой части $p(w \mid d) = n_{dw}/n_d$ найти неизвестные условные распределения в правой части: $p(w \mid t)$ для каждой темы $t \in T$ и $p(t \mid d)$ для каждого документа $d \in D$, а также определить оптимальное число тем $|T|$.

Большинство тематических моделей [8, 9, 10, 11] оценивают вероятности тем $p(t \mid d, w)$ для каждого слова w в каждом документе d. Зная эти вероятности, возможно оценить число троек:

$n_{dwt} = n_{dw} p(t \mid d, w)$ — в которых термин w документа d связан с темой t,
$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — в которых термин документа d связан с темой t,
$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — в которых термин w связан с темой t,
$n_t = \sum\limits_{d \in D} \sum\limits_{w \in W} n_{dwt}$ — связанных с темой t,

и затем по ним найти частотные оценки искомых условных вероятностей:

$$\hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(w \mid d, t) = \frac{n_{dwt}}{n_{dt}}. \tag{5}$$

Чтобы оценить качество тематической модели, необходимо проверить, выполняется ли гипотеза условной независимости (3) — важнейшее базовое предположение модели (4) — для каждой пары документ–тема (d, t). Тема t описывается распределением $\hat{p}(w \mid t)$. Выборка слов документа d, относящихся к теме t, согласно модели, образует эмпирическое распределение $\hat{p}(w \mid d, t)$. Оба распределения оцениваются согласно (5) в процессе построения тематической модели. Чтобы проверить, действительно ли данная выборка могла быть получена из распределения $\hat{p}(w \mid t)$, воспользуемся критерием согласия, основанным на статистике хи-квадрат (1):

$$X^2_{dt} = n_{dt} \sum_{w:\, n_{wt} > 0} \frac{\big(\hat{p}(w \mid d, t) - \hat{p}(w \mid t)\big)^2}{\hat{p}(w \mid t)}. \tag{6}$$

Число различных слов в теме может быть намного больше, чем число слов в документе. Следовательно, мы имеем дело с разреженными распределениями, к которым неприменим асимптотический критерий хи-квадрат. Поэтому будем строить статистические тесты методом сэмплирования, для каждой темы $t \in T$ отдельно.

Экспериментально установлено, что для больших корпусов текстов на естественных языках закон Ципфа или более сложные параметрические законы (например Ципфа–Мандельброта) выполняются с неплохой точностью [3, 4]. Для ускорения проверки гипотезы условной независимости предлагается двухэтапный тест. Сначала проверяется согласие каждой темы t с выбранным параметрическим законом. Если согласие есть, то строится один регрессионный тест для всех таких тем. Для каждой из остальных тем строится отдельный тест на основе сэмплирования.

Сэмплирование без возвращений

Проверки согласия документных эмпирических распределений $\hat{p}(w \,|\, d, t)$, $d \in D$ с распределением $\hat{p}(w \,|\, t)$, вообще говоря, не являются независимыми, поскольку имеется тождество, связывающие эти распределения друг с другом:

$$\hat{p}(w \,|\, t) = \sum_{d \in D} \hat{p}(w \,|\, d, t) \hat{p}(d \,|\, t). \tag{7}$$

Документы являются выборками без возвращений из распределения $\hat{p}(w \,|\, t)$, тогда как обычно критерии согласия предполагают выборку с возвращениями. Наличие дополнительного ограничения (7) может и не влиять на результаты тестов или влиять несущественно, особенно на коллекциях большого размера. Однако это лишь предположение, которое необходимо проверить. Для этого построим более точный тест на основе сэмплирования *без возвращений*, учитывающий, что последовательность слов, образующих тему t, разрезается на документы в пропорциях $\hat{p}(d \,|\, t)$.

Построение теста сэмплированием без возвращений. Возьмем последовательность терминов длины n_t, образующую распределение $\hat{p}(w \,|\, t)$. Сгенерируем N случайных перестановок этой последовательности. Разрежем каждую из полученных последовательностей W_j, $j = 1, \ldots, N$ на «документы» — подпоследовательности терминов W_{jd} длины n_{dt} каждая, $d \in D$. По каждому «документу» W_{jd} построим эмпирическое распределение $\hat{p}_j(w \,|\, d, t)$ и вычислим значение статистики хи-квадрат X^2_{jd}. Для каждого $d \in D$ по множеству значений статистики $X^2_{1d}, \ldots, X^2_{Nd}$ построим эмпирическую функцию распределения $\hat{F}_d(X^2)$ и вычислим её $(1-\alpha)$-квантиль $\hat{F}_{d,1-\alpha}$. Число N рекомендуется брать не менее 1000 при типичном значении $\alpha = 0.05$.

Отметим, что в тесте без возвращений квантиль строится для каждого документа d, тогда как тест с возвращениями строился для каждого значения длины документа n. Построение теста без возвращений более ресурсоемко и требует $O(n_t N \log N)$ операций вместо $O(n_{\max} N \log N)$, где $n_{\max} = \max\limits_{d \in D} n_{td}$.

Применение теста сэмплированием без возвращений. Проверка гипотезы условной независимости для пары документ–тема (d, t) заключается в вычислении статистики X^2_{dt} по формуле (6) и проверке неравенства $X^2_{dt} > \hat{F}_{d,1-\alpha}$. Если оно выполнено, то гипотеза условной независимости отвергается для данной пары (d, t).

Вычислительные эксперименты

Эксперименты проводились на коллекции из $|D| = 2000$ авторефератов диссертаций на русском языке. Мощность словаря после предварительной обработки данных (лемматизации и удаления стоп-слов) составляет $|W| = 20211$ слов, длина документов от 1000 до 4000 слов. Строились две тематические модели — PLSA [8] и LDA-GS [9, 10] с помощью алгоритма, описанного в [12]. Число тем $|T| = 100$.

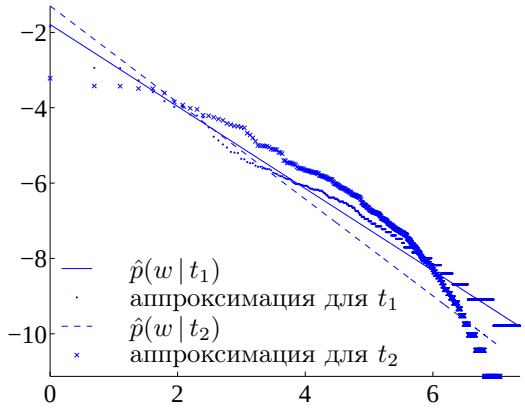


Рис. 8: Аппроксимация эмпирических распределений слов законом Ципфа (для двух тем, в логарифмических осях).

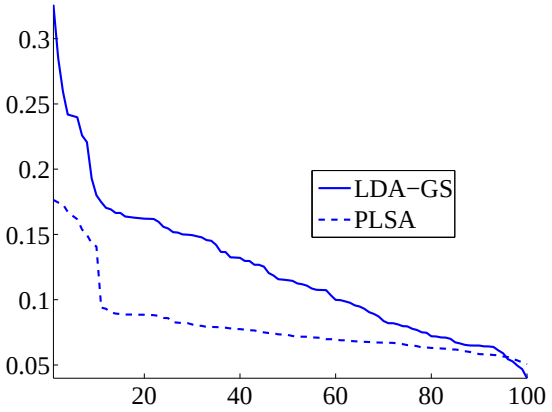


Рис. 9: Доля документов, для которых гипотеза условной независимости отклоняется (в порядке убывания).

Выполняется ли закон Ципфа для тем? На рис. 8 показаны графики эмпирических распределений и закона Ципфа для двух из 100 тем t_1 и t_2 в модели LDA, в логарифмических осях. По горизонтальной оси откладывается логарифм номера слова, слова упорядочены по частоте. По вертикальной оси откладывается логарифм вероятности слова. Оптимальные значения параметра закона Ципфа: $s = 1.04$ для t_1, $s = 1.28$ для t_2. Хотя «на глаз» соответствие неплохое, особенно для t_1, нулевая гипотеза отклоняется для обоих тем. Более того, большинство тем согласуются с законом Ципфа лишь при крайне низких уровнях значимости, меньших 0.05. Это объясняется тем, что при выборках длины n_t порядка 10^3–10^5 критерии согласия чувствительны даже к незначительным различиям распределений, и одного параметра в законе Ципфа не достаточно для описания эмпирических распределений.

Сравнение тестов без возвращений и с возвращениями.

Для модели PLSA рассматривается одна тема из $|D_t| = 1992$ документов суммарной длины $n_t = 87026$ слов. В тестах без возвращений и с возвращениями нулевая гипотеза принимается для 1674 и 1688 документов соответственно. Решения отличаются на 22 документах из 1992. Оба теста дают примерно одинаковый результат: гипотеза условной независимости отклоняется для 15% документов.

Для модели LDA-GS рассматривается тема из $|D_t| = 1114$ документов суммарной длины $n_t = 63805$ слов. Нулевая гипотеза принимается для 1032 и 1035 документов соответственно. Решения отличаются на 7 документах из 1114. Оба теста снова дают примерно одинаковый результат: нулевая гипотеза отклоняется для 7% документов.

Таким образом, результаты тестов без возвращений и с возвращениями почти одинаковы, однако тест с возвращениями менее ресурсоемкий.

На рис. 9 показан результат сравнения моделей PLSA и LDA по всем темам. По вертикальной оси откладывается доля документов, для которых отклоняется гипотеза условной независимости. По горизонтальной оси откладываются темы в порядке убывания долей (порядки тем для двух моделей, естественно, не совпадают). Модель LDA строит темы менее аккуратно, что, возможно, объясняется применением смещенных (сглаженных) частотных оценок условных вероятностей в LDA, в то время как PLSA основан на несмещенных оценках максимального правдоподобия. Однако в обеих моделях доля документов, не прошедших тест, превышает уровень значимости 0.05 для всех тем. Это может быть объяснено выбором неоптимального (заниженного) числа тем $|T| = 100$. Более полный статистический анализ качества тематических моделей PLSA и LDA выходит за рамки данной работы.

Выводы

Предложены критерии согласия на основе сэмплирования для разреженных дискретных распределений, выходящих за границы применимости классических асимптотических критериев. Предложен рекуррентный алгоритм построения теста на основе сэмплирования. Для параметрического случая, когда проверяется согласие эмпирических данных с распределением Ципфа, построен регрессионный тест, подобрана модель регрессии и проведен анализ ошибок первого и второго рода. Рассмотрено применение предложенных тестов для проверки гипотезы условной независимости — ключевого предположения вероятностных тематических моделей коллекций текстовых документов. Экспериментально показано, что в случае большого числа документов нет необходимости строить точный тест без возвращений, и можно пользоваться вычислительно более эффективным тестом с возвращениями.

Работа выполнена при поддержке Министерства образования и науки Российской Федерации (Государственный контракт 07.524.11.4002).

Литература

[1] *Zelterman D.* Goodness-of-fit tests for large sparse multinomial distributions // *Journal of the American Statistical Association.* 1987. - Vol. 398. No. 82. P. 624–629.

[2] *von Davier M.* Bootstrapping goodness-of-fit statistics for sparse categorical data — results of a monte carlo study // *Methods of Psychological Research Online.* 1997. Vol. 2. No. 2.

[3] *Бриллюэн Л.* Наука и теория информации. — М.: Государственное издательство физико-математической литературы, 1960. — 391 с.

[4] *Gelbukh A., Sidorov G.* Zipf and heaps laws' coefficients depend on language // Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science. — Springer-Verlag, 2001. P. 332–335.

[5] *Strijov V.* Search for a parametric regression model in an inductive-generated set // *Computational technologies.* 2007. - Vol. 12. No. 1. P. 93–102.

[6] *Strijov V.* MVR Composer. — 2012. http://strijov.com/?p=84.

[7] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of Computer Science in China.* 2010. Vol. 4. No. 2. P. 280–301.

[8] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. P. 50–57.

[9] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* 2003. Vol. 3. P. 993–1022.

[10] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* 2004. Vol. 101. No. Suppl. 1. P. 5228–5235.

[11] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. 2009.

[12] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование.* 2012. Т. 4. № 4. — С. 693–706.

Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным*

Вальков А. С.[1]***, Кожанов Е. М.***[2]***, Медведникова М. М.***[3]***, Хусаинов Ф. И.***[4]
valkov@forecsys.ru, vinger4@gmail.com, medvmasha@rambler.ru, f-husainov@yandex.ru
1 — Вычислительный центр РАН, 2 — Московский государственный технический университет имени Н. Э. Баумана, 3 — Московский физико-технический институт, 4 — Российская открытая академия транспорта Московского государственного университета путей сообщения (МИИТ)

Предложен алгоритм непараметрического прогнозирования загруженности железнодорожных узлов РЖД по историческим данным. Алгоритм основан на свертке эмпирической плотности распределения значений временного ряда с функцией потерь. В работе исследуются свойства авторегрессионной прогностической модели. Алгоритм проиллюстрирован данными загруженности железнодорожных узлов Омской области за 2007 и 2008 гг.

Ключевые слова: *временные ряды, прогнозирование, загрузка железнодорожного узла, непараметрический метод, эмпирическое распределение.*

Nonparametric forecasting of railroad stations occupancy according to historical data*

Valkov A. S.[1]***, Kozhanov E. M.***[2]***, Medvednikova M. M.***[3]***, Husainov F. I.***[4]
1 — Computing Center of the Russian Academy of Sciences; 2 — Bauman Moscow State Technical University; 3 — Moscow Institute of Physics and Technology; 4 — Moscow State University of Railway Engineering

The authors propose a method of nonparametric forecasting of railroad stations occupancy according to historical data. The algorithm is based on convolution of empirical density of distribution of time series values and loss function. The features of autoregressive prognostic model are investigated. The algorithm is illustrated by railroad stations occupancy data in Omsk region in 2007 and 2008.

Keywords: *time series, forecasting, railroad station occupancy, non-parametric method, empirical distribution.*

Введение

Прогнозирование потребностей в вагонах у заказчиков РЖД в узлах погрузки/разгрузки с учетом временных интервалов доставки, а также использование загруженности железнодорожных узлов является проблемой, которую необходимо решить для повышения эффективности транспортировки грузов. Данная работа посвящена решению задачи прогнозирования загруженности железнодорожных узлов. Прогноз выполняется на основании исторических знаний о приходящих на станцию и уходящих со станции вагонах. При этом в качестве единицы учета рассматривается блок вагонов, неделимый на всем протяжении маршрута.

Работа выполнена при финансовой поддержке РФФИ, проект № 11-07-13154.

Используется формат данных, образец которых представлен в табл. 1. Каждая строка (запись в базе данных) содержит информацию о дате погрузки, станции отправления, станции назначения, количестве вагонов, которые прошли по маршруту от станции отправления до станции назначения, коде груза, роде вагонов, суммарном весе груза и признаке маршрутной отправки.

Согласно используемым данным железнодорожный узел рассматривается без детализации по путям и по очередности отправления блоков вагонов. Для прогноза не используются внешние данные.

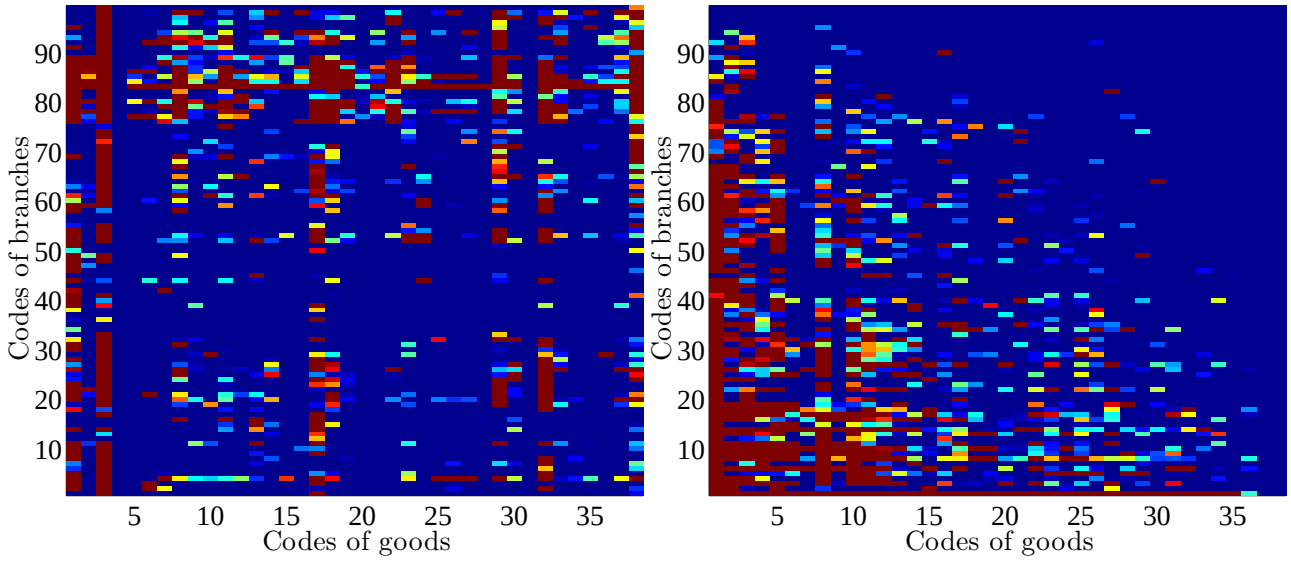
Для решения задачи требуется сформировать прогноз отправления/погрузки грузов в заданном периоде:

1) на месяц посуточно;
2) на месяц подекадно;
3) на квартал помесячно;
4) на год помесячно;
5) на год поквартально;
6) на период больше года;

и прогноз отправления/погрузки грузов с разложением:

1) по группам грузов;
2) по родам подвижного состава;
3) по комбинированному разложению, учитывающему перечисленные варианты.

Для получения адекватного прогноза должна быть решена задача определения рационального уровня детализации прогноза (по станции или по группе станций). Существенные внешние ограничения возникают в связи с тем, что в одном составе перемещаются вагоны, принадлежащие различным собственникам, и с тем, что возникают запреты на движение товарных поездов из-за необходимости обеспечения возможности высокоскоростного передвижения.



(a) Без сортировки (b) С сортировкой

Рис. 1: Суммарная матрица перевозок

Используемый в данной работе алгоритм основан на алгоритме квантильной регрессии [1, 2], модифицированным сверткой гистограммы с функцией потерь. Главное преимущество такого подхода заключается в возможности учета стоимости ошибки прогнозирования в прикладной задаче.

Основные методы непараметрической регрессии, такие как ядерное сглаживание, сглаживание сплайнами, авторегрессия, скользящее среднее и др., описаны в [3, 5, 4, 6]. Они заключаются в присвоении имеющимся значениям временного ряда некоторых весов и комбинации взвешенных значений для получения прогноза. Также для решения подобных задач применяют нейронные сети [7, 8].

Для построения прогностической модели предлагается использовать непараметрические методы прогнозирования. В частности, предполагая временной ряд локально-стационарным (выполнено условие Дики-Фуллера [6]), предлагается построить гистограмму распределения его значений и вычислить свертку гистограммы с экспертно заданной функцией потерь для каждого возможного прогнозируемого значения. Оптимальным прогнозом является то значение центра сегмента гистограммы, которое доставляет минимальное значение свертки. Также проверяется применимость на практике данной модели к прогнозированию нестационарных временных рядов.

В качестве базового алгоритма для сравнения полученных прогнозов используется модель авторегрессионного скользящего среднего ARMA [9, 10, 11].

Алгоритм непараметрического прогнозирования временного ряда

Задан временной ряд $\mathbf{x} = \{x_i\}_{i=1}^{T}$ и горизонт отсрочки прогноза h (число отсчетов от конца временного ряда до точки прогноза, включительно). Требуется спрогнозировать следующую точку временного ряда $\mathbf{x}$ так, чтобы выполнялось условие оптимальности свертки гистограммы, построенной по значениям временного ряда, и функции потерь. Предполагается, что ряд стационарен, иными словами, распределение прогнозируемого значения в точке равно распределению точек заданного временного ряда.

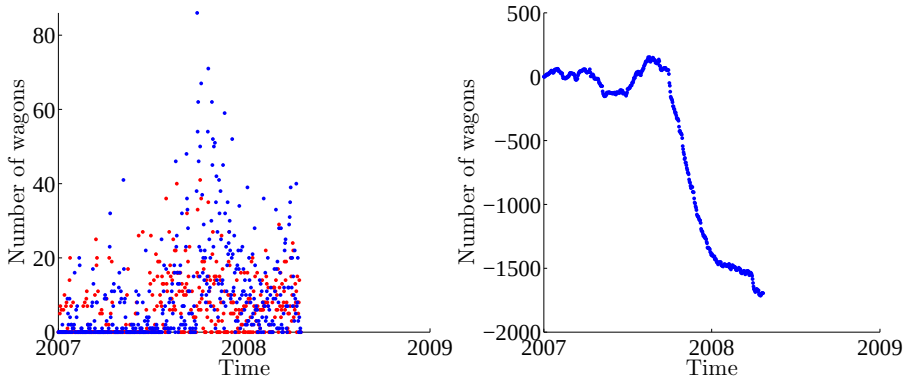
Для построения прогностической модели используются элементы квантильной регрессии. По временному ряду $\mathbf{x}$ построим гистограмму $\mathcal{H}$ — набор пар

$$\mathcal{H} = \{(y_k, g_k)\}_{k=1}^{K}, \tag{1}$$

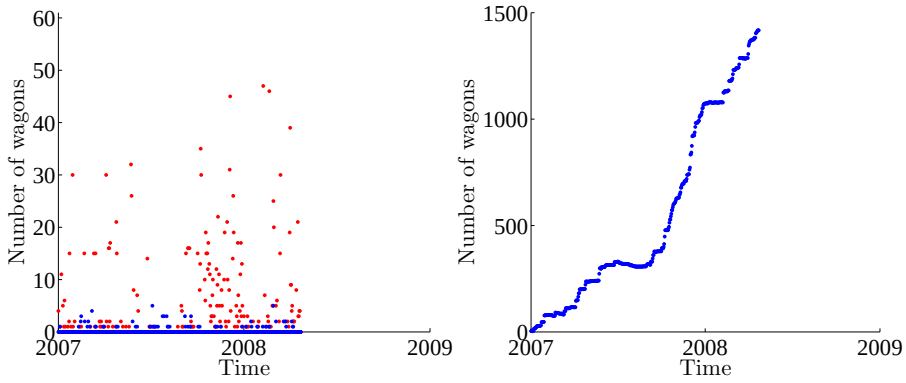
где K — число интервалов $[y_k^{\min}, y_k^{\max}]$ со средним значением y_k, на которые разбита ось значений ряда, g_k — высота столбца гистограммы на интервале y_k, которая равна взвешенной сумме количества точек ряда, попавших в этот интервал. В алгоритме квантильной регрессии [1, 2] для прогноза используется значение y_k, соответствующее самому высокому столбцу (моде) гистограммы. В данной работе предлагается модификация алгоритма квантильной регрессии.

Введем функцию потерь $L(\hat{y}, y)$ — штраф за несоответствие прогнозируемого значения $\hat{y}$ историческому значению y. Далее будем использовать одну из трех функций потерь:

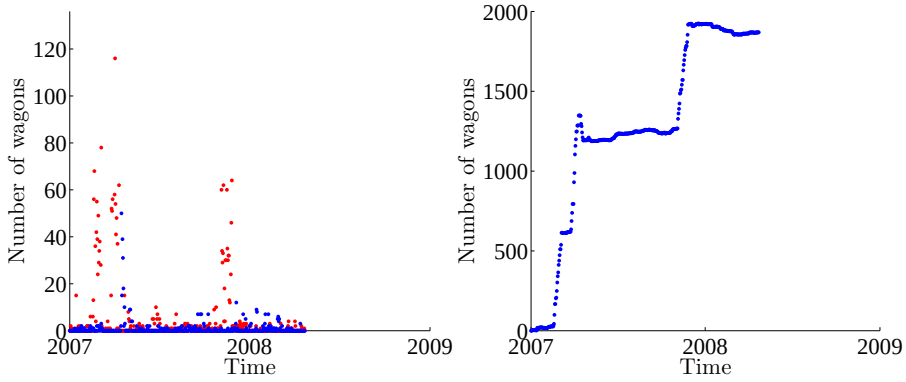
1) $L(z, x) = (z - x)^2$;
2) $L(z, x) = |z - x|$;
3) $L(z, x) = \begin{cases} 0, & \text{если } |z - x| < a; \\ |z - x| - a, & \text{если } |z - a| \geqslant a, \text{ где } a > 0 \text{ — экспертно заданный параметр.} \end{cases}$



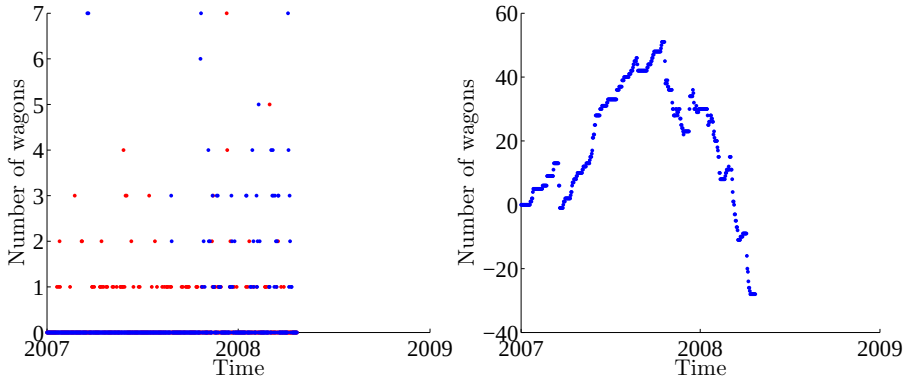
(a) Ветка 0 (b) Ветка 0



(c) Ветка 6 (d) Ветка 6



(e) Ветка 24 (f) Ветка 24



(g) Ветка 75 (h) Ветка 75

Рис. 2: Прибывшие (красный) и отправленные (синий) вагоны и число вагонов на ветке

Построение гистограммы. Для каждой точки i временного ряда $\mathbf{x}$ определим ее вес как произведение $w_i = w_i^F w_i^H$. Сомножитель w_i^F задает показательную весовую функцию

$$w_i^F = v^{\frac{-i+T+h}{F}} \in (0, 1], \tag{2}$$

убывающую к началу временного ряда и равную 1 в точке прогноза. Сомножитель w_i^H задан как

$$w_i^H = \begin{cases} K(i_H, PH), & \text{если } H > 0; \\ 1, & \text{если } H = 0, \end{cases} \tag{3}$$

где индекс H вычисляется в результате решения оптимизационной задачи

$$i_H = \min_{n=0,\ldots,\mathrm{floor}\left(\frac{T+h}{P}\right)} |T + H - nP - i|. \tag{4}$$

Эта формула задает вес i-той точки, соответствующий годовой сезонности. Ядро задается выражением:

$$K(x, z) = \begin{cases} \left(1 - \left(\frac{x}{z}\right)^2\right)^2, & \text{если } |x| < z; \\ 0, & \text{иначе.} \end{cases}$$

Взвешенные точки $x_i w_i$ используются для построения гистограммы $\mathcal{H}$ (1).

Настраиваемые параметры: $v \in [0, 1]$ в выражении (2) — параметр показательного взвешивания точек ряда, параметр «забывания»; $H \in [0, 0.5]$ в выражениях (3) и (4) — параметр ядра весовой функции для годовой сезонности, половина ширины «шапки» годовой сезонности.

Ненастраиваемые параметры: P в выражениях (3) и (4) — длина годового сезонного периода (обычно $P = 365$); $w^{\min}$ в выражении (5) — минимальный допустимый вес; F в выражении (2) — нормировочная константа «забывания». Предлагается выбрать F следующим образом $F = (T + H)\varepsilon \log_{10}(0.1), \quad \varepsilon = 10^{-3}$.

Выберем границы гистограммы, число столбцов и разбиение на столбцы следующим образом:

1) пусть n — число точек x_i, для которых $w_i > w^{\min}$;
2) выберем число столбцов (обоснование см. в [12]) $K = \lceil 3\sqrt[3]{n}\,\rceil$, если $K < 5$, то $K = 5$, если $K > 100$, то $K = 100$;
3) границы $y_1 = \min\limits_{i:w_i > w^{\min}}(x_i)$, $y_k = \max\limits_{i:w_i > w^{\min}}(x_i)$;
4) столбцы выбираются равной ширины.

Для каждого $k = 1, \ldots, K$ высота столбца гистограммы g_k равна

$$g_k = \sum_{i=1}^{T} w_i [x_i \in y_k][w_i > w^{\min}], \tag{5}$$

где выражение $[\cdot]$ равно 1, если в скобках стоит истинное логическое выражение, и 0 — в противном случае.

Алгоритм непараметрического прогнозирования. Полученная гистограмма $\mathcal{H}$ используется для построения прогноза. Прогнозируемое значение ряда x_{T+h} находится как

значение $\hat{y} \in \{y_1, \ldots, y_K\}$, соответсвующее оптимальному значению свертки распределения $\{g_k\}_{k=1}^K$ и функции потерь L:

$$\hat{y} = \underset{z \in \{y_1, \ldots, y_K\}}{\arg\min} \sum_{k=1}^{K} g_k L(z, y_k). \tag{6}$$

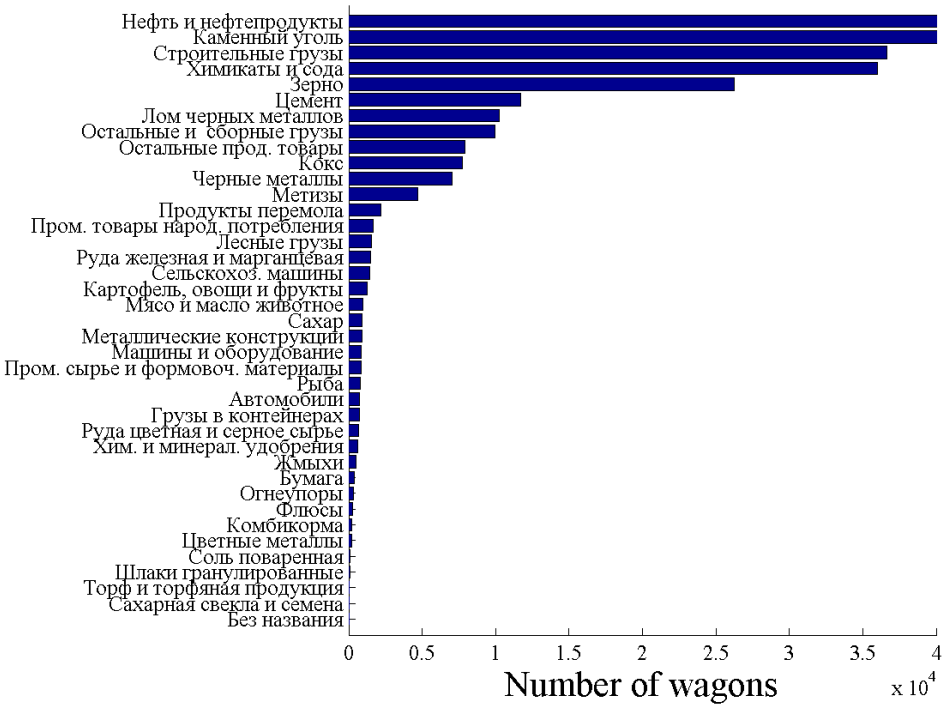


Рис. 3: Вагоны с разными типами грузов

Тест Дики-Фуллера на стационарность временного ряда является одним из тестов на единичные корни. Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд $I(0)$:

$$\Delta x_i = x_i - x_{i-1} \sim I(0).$$

При помощи этого теста проверяют значение коэффициента a в авторегрессионном уравнении первого порядка AR(1):

$$x_i = a x_{i-1} + \varepsilon_i,$$

где ε — ошибка. Если $a = 1$, то процесс имеет единичный корень и ряд $\mathbf{x}$ не стационарен. Если $|a| < 1$, то ряд стационарный. Приведенное авторегрессионное уравнение можно переписать в виде

$$\Delta x_i = b x_{i-1} + \varepsilon_i,$$

где $b = a - 1$. Поэтому проверка гипотезы о единичном корне в данном представлении означает проверку нулевой гипотезы $b = 0$ против альтернативы $b < 0$. Статистика теста (DF-статистика) — t-статистика для проверки значимости коэффициентов линейной регрессии $y = bx$:

$$t = \frac{\hat{b} - b}{S_b},$$

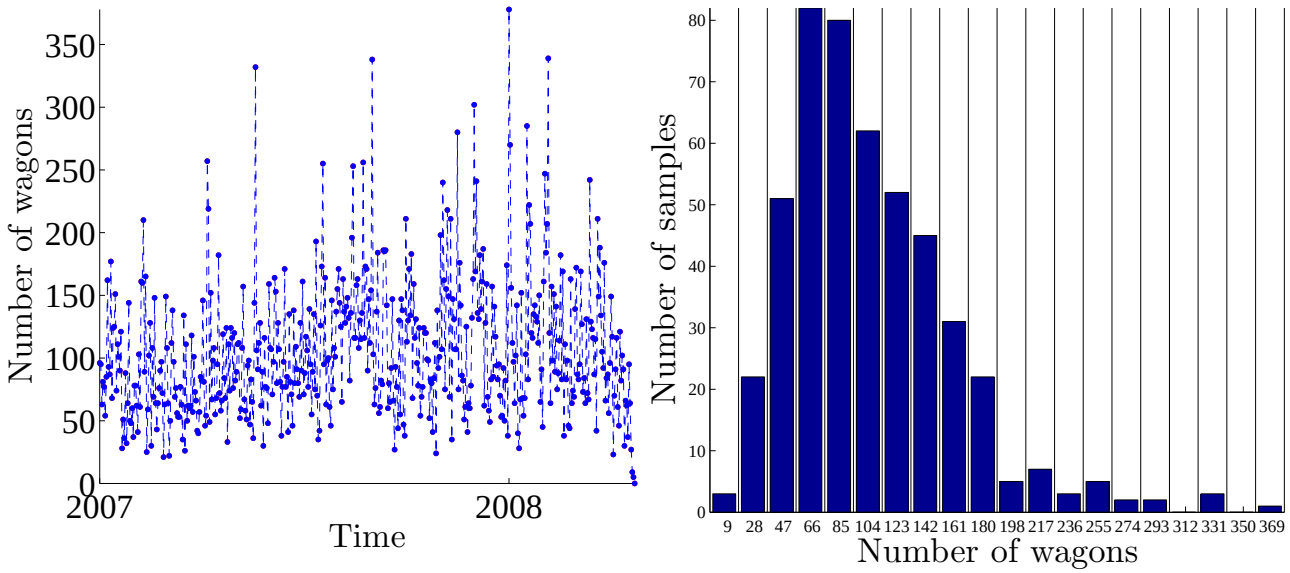
где $\hat{b}$ — оценка коэффициента регресии по выборке и

$$S_b = \frac{S}{S_x\sqrt{n-1}}; \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{b}x_i)^2;$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i.$$

Распределение DF-статистики выражается через винеровский процесс и называется распределением Дики-Фуллера [6].

Считаем выполнение этого теста необходимым для применения прогностической модели.



(a) Временной ряд (b) Гистограмма

Рис. 4: Прибытие вагонов с нефтью и нефтепродуктами

Базовый алгоритм прогнозирования

В качестве базового алгоритма для проверки полученных результатов используется авторегрессионное скользящее среднее (ARMA) [9, 10, 11]. Пусть, как и ранее, задан временной ряд $\mathbf{x} = \{x_i\}_{i=1}^{T}$. Авторегрессионная модель (AR):

$$x_i = c + \sum_{\tau=1}^{p}\varphi_\tau x_{i-\tau} + \varepsilon_i,$$

где $\varphi_1, \ldots, \varphi_p$ — параметры модели, c — константа, ε_i — шум. Модель скользящего среднего (MA):

$$x_i = \mu + \varepsilon_i + \sum_{\tau=1}^{q}\theta_\tau\varepsilon_{i-\tau},$$

где $\theta_1, \ldots, \theta_q$ — параметры модели, μ — математическое ожидание x_i, $\varepsilon_i, \varepsilon_{i-1}, \ldots$ — шумы. Модель ARMA:

$$x_i = c + \varepsilon_i + \sum_{\tau=1}^{p}\varphi_\tau x_{i-\tau} + \sum_{j=1}^{q}\theta_j\varepsilon_{i-j}.$$

Шумы ε_i обычно принимают независимыми одинаково нормально распределенными случайными величинами с нулевым математическим ожиданием: $\varepsilon_t \in \mathcal{N}(0, \sigma^2)$.

Оптимизация параметров авторегрессионной модели описана в [11].

Ретроспективный прогноз и оценка качества прогностической модели

В ходе вычислительного эксперимента составлялся прогноз прибытия и отправления вагонов с различными типами грузов на день, неделю и месяц по всем железнодорожным веткам. В алгоритме непараметрического прогнозирования для построения гистограммы использовались H точек, предшествующие точке (интервалу) прогноза. Окно в H точек перемещалось по временному ряду с шагом в одну точку с построением прогноза для каждого шага.

Для оценки качества прогностической модели использовалась средняя доля ошибки:

$$MAPE = \frac{\sum\limits_{i=1}^{h} \frac{|\hat{y}-y_i|}{y_i}}{h}, \tag{7}$$

где $\hat{y}$ — полученное при прогнозе значение, h — горизонт отсрочки прогноза. Затем ошибка прогнозирования усреднялась по всем интервалам. Полученные результаты сравнивались с результатами модели ARMA.

Вычислительный эксперимент

Входные данные. В эксперименте использованы данные о посуточной загруженности железнодорожных узлов РЖД с 1 января 2007 года по 22 апреля 2008 года. В табл. 1 приведен пример записи.

Таблица 1: Вид записи базы данных железнодорожных перевозок

Дата погрузки	Станция отправления	Станция назначения	Количество вагонов	Код груза	Род вагона	Суммарный вес груза	Признак маршрутной отправки
2007-01-01	020108	932902	1	1	216	56	9

Коды станций представляют собой шестизначные числа. Станции, в коде которых две первые цифры совпадают, входят в одну железнодорожную ветку. Станций отпраления 1566, станций назначения 1902, веток 99. Код груза — натуральное число от 1 до 43; также имеются перевозки, где код груза не указан. Род вагона — натуральное число, в имеющихся данных 75 различных типов вагонов. Поскольку имеющихся данных недостаточно, для того чтобы проследить годовую периодичность временных рядов, то в ходе эксперимента наличие периодичности не учитывалось.

На рис. 1 изображена матрица перевозок. По горизонтали отложены коды грузов, по вертикали — номера железнодорожных веток. Цвет каждой ячейки соответствует числу вагонов с данным типом груза, прошедших за все время наблюдения через данную ветку. Большим значениям соответствуют ячейки красных оттенков, малым — синих оттенков. На рис. 1(a) показана матрица с неотсортированными столбцами и строками. Коды грузов и коды веток совпадают с отмеченными на осях числами. На рис. 1(b) строки и столбцы отсортированы по убыванию суммы их элементов (суммируются значения в столбцах,

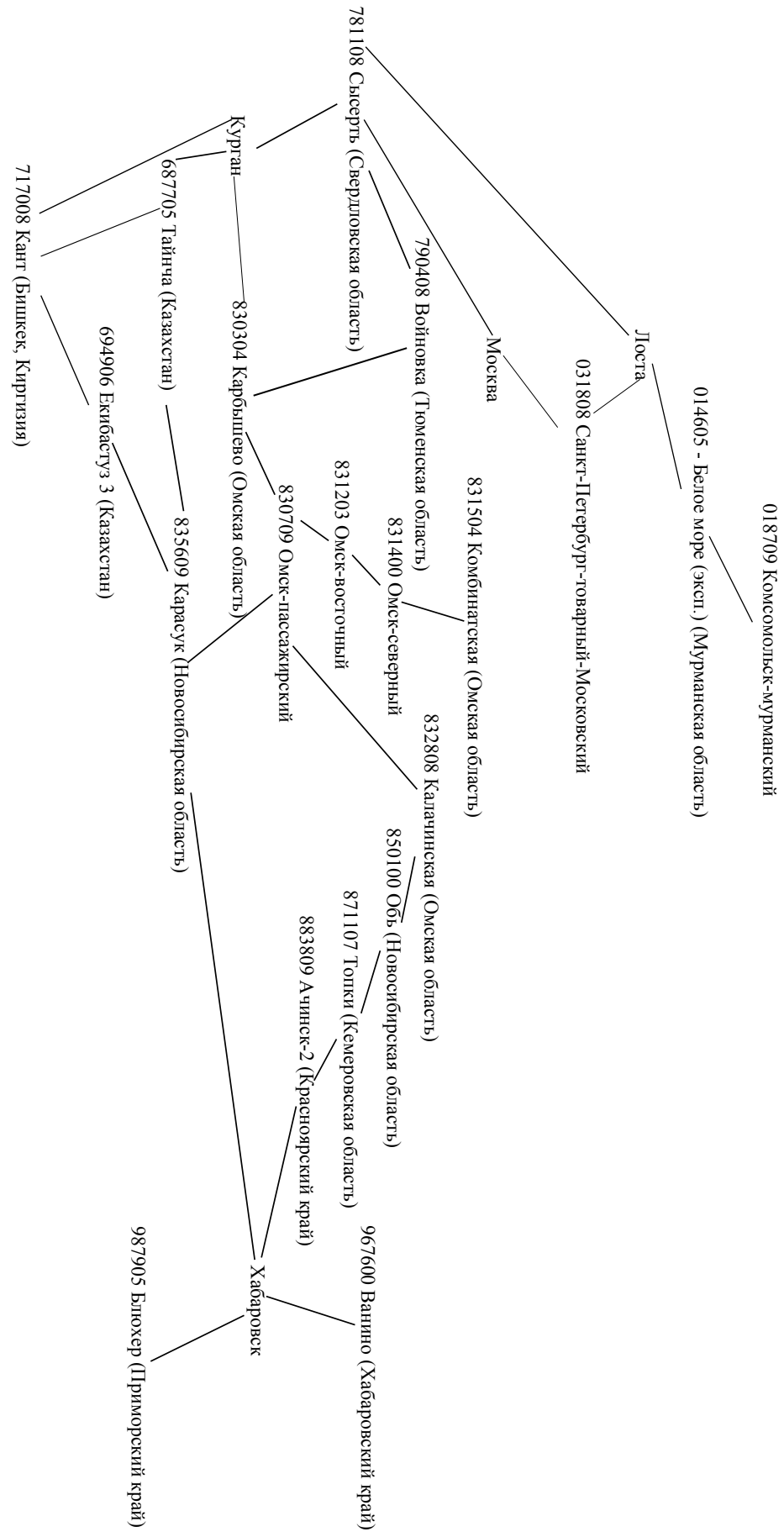
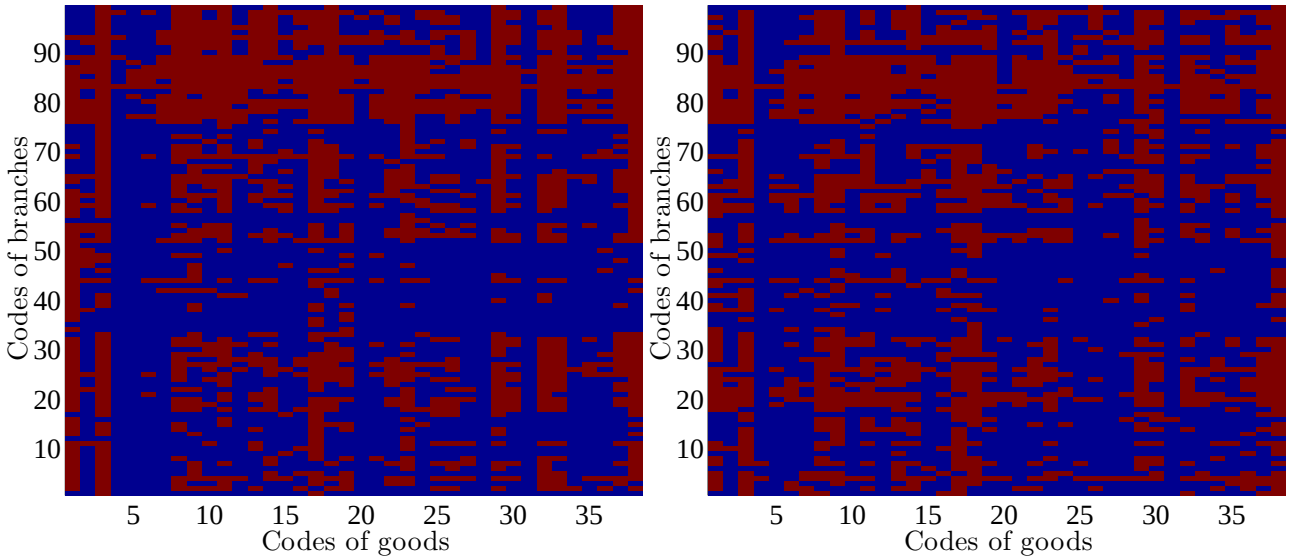
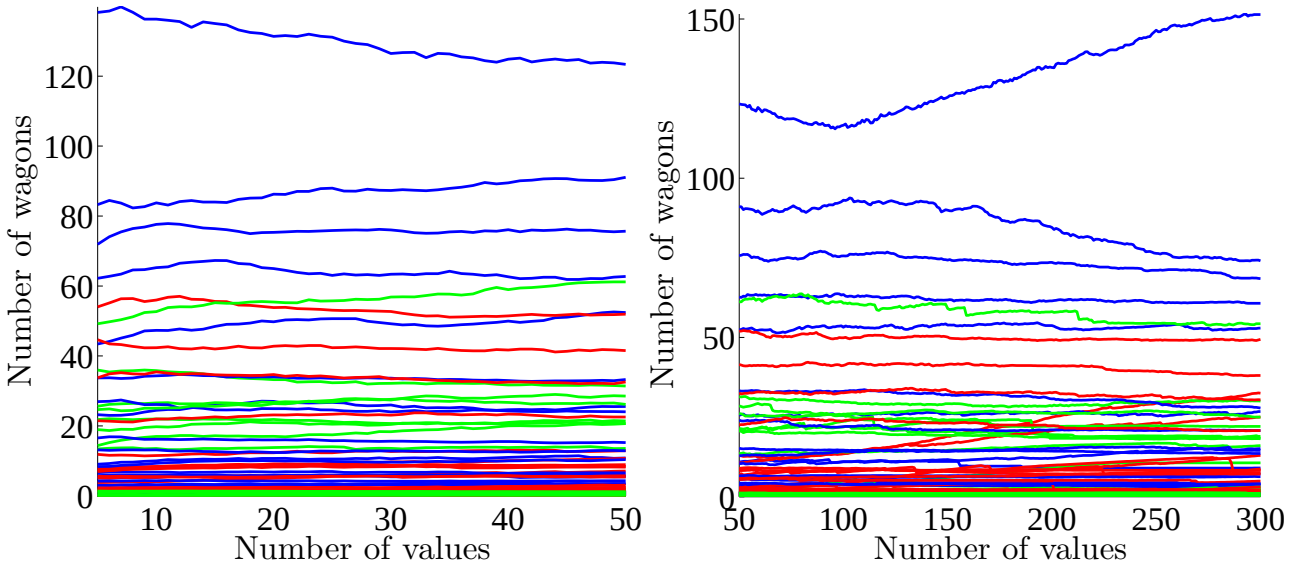


Рис. 5: Граф станций



(a) Прибытие (b) Отправление

Рис. 6: Тест Дики-Фуллера для временных рядов

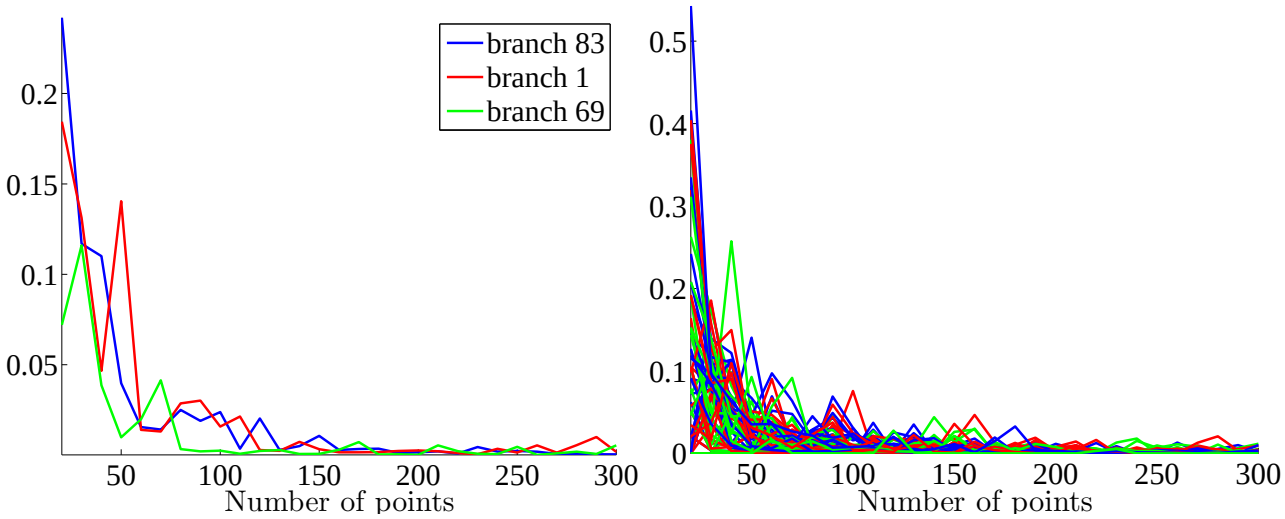


(a) Малое количество точек (b) Большое количество точек

Рис. 7: Ошибки прогнозирования

затем суммы сортируются по убыванию и переставляются столбцы; затем та же операция проводится со строками), и коды грузов и веток не совпадают с числами на осях.

Рис. 2 показывает посуточное перемещение всех типов вагонов по четырем веткам. По оси абсцисс отложены даты, по оси ординат — количество вагонов. На графиках в левом столбце красными точками отмечено количество прибывших на ветку вагонов в течение суток, синими — число отправленных с ветки за тот же период. В правом столбце изображено количество оставшихся на ветке в течение суток вагонов в предположении, что изначально на ветке вагонов не было (этим объясняется возможность отрицательного числа вагонов). Графики показывают, что через разные ветки проходит различное число вагонов. Динамика числа вагонов на разных ветках также различна. Их число может возрастать, убывать или не иметь постоянного тренда. Причем резкие скачки на графике справа соответствуют пикам синего или красного цвета на графике слева, в зависимости от того, уменьшается или увеличивается количество вагонов.



(a) Ветки 1, 69, 83 (b) Все ветки

Рис. 8: Стабилизация гистограммы, прибытие вагонов с нефтью

На рис. 3 изображена гистограмма числа вагонов с разными типами грузов, прошедших через все станции полигона за все время наблюдения. По вертикали отмечены названия грузов, по горизонтали — количество вагонов. Самые большие столбцы, соответствующие перевозкам нефти и нефтепродуктов и каменного угля обрезаны на значении 40 000, чтобы более короткие столбцы были видны на диаграмме.

На рис. 4 изображен временной ряд и гистограмма прибытия на 83 ветку вагонов с нефтью и нефтепродуктами. По оси абсцисс гистограммы отложено число вагонов, по оси ординат — количество наблюдений, соответствующее числу вагонов в интервале. По оси абсцисс временного ряда отложена дата, по оси ординат — число вагонов, пришедших на ветку за сутки.

На рис. 5 показана топология самых загруженных станций. Цифры обозначают коды станций, между соединенными станциями есть сообщение.

Поскольку стационарность временных рядов считается необходимой для использования рассматриваемой прогностической модели, для всех рядов был проведен тест на стационарность. На рис. 6 показаны результаты теста Дики-Фуллера для временных рядов

для каждой ветки и каждого типа груза. Красным обозначены ряды, не прошедшие тест, синим — прошедшие. Использовалась реализация теста в среде MatLab. В ходе эксперимента предлагаемая в данной работе непараметрическая модель использовалась для прогноза всех рядов с целью проверки ее применимости на практике к нестационарным рядам.

Выбор параметров модели. Для построения прогностической модели выбиралась длина предыстории, используемая для построения гистограммы, и функция потерь для свертки.

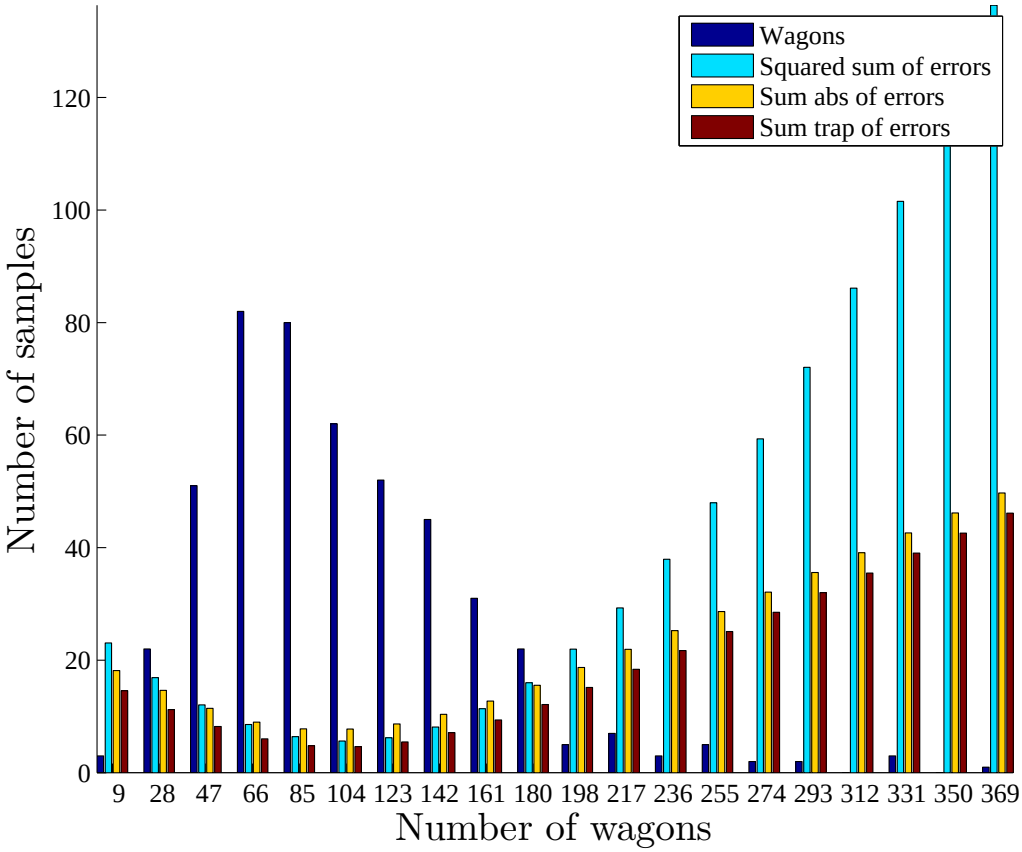


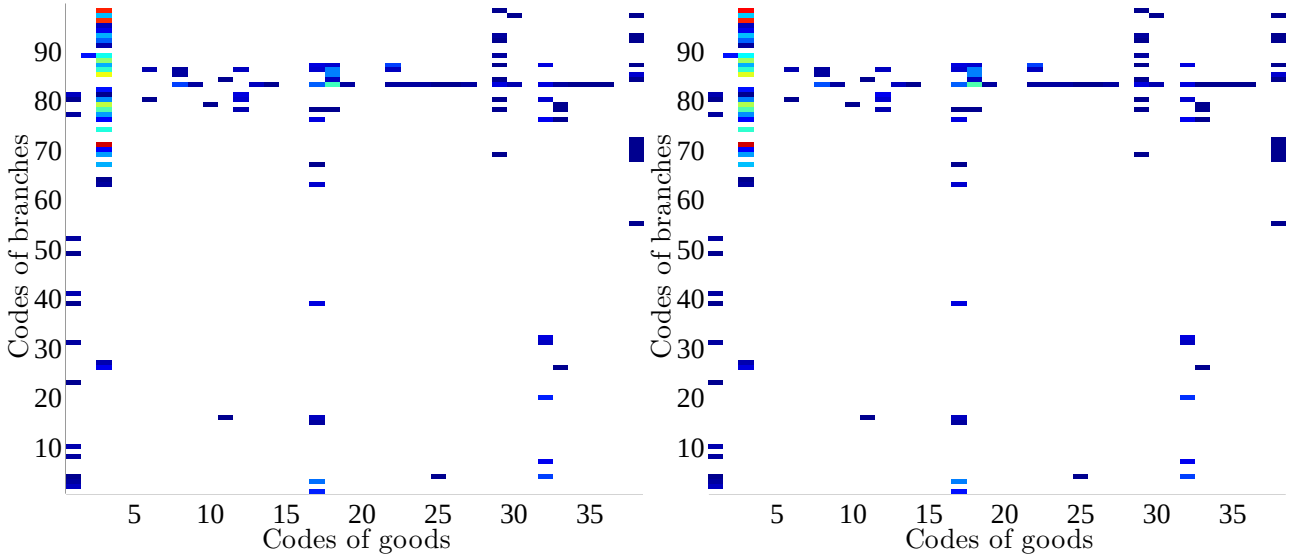
Рис. 9: Свертка гитограммы с различными функциями потерь

Для выбора длины предыстории были использованы два способа. В первую очередь исследовалась зависимость средней ошибки прогнозирования от длины предыстории. Результаты исследования представлены на рис. 7. По оси абсцисс графиков отложено количество точек, использованных для построения гистограммы, по оси ординат — средняя ошибка в количестве вагонов. Прогноз был сделан на один день для точек с номерами $301, \ldots, 478$. На графиках показано среднее по всем прогнозам значение модуля отклонения от реального значения ряда (в количестве вагонов) в зависимости от количества точек, использованных при построении гистограммы. Число столбцов гистограммы фиксируется равным оптимальному числу столбцов для гистограммы, построенной по всему временному ряду (см. описание алгоритма построения гистограммы). Использована квадратичная функция потерь. Эксперимент проводился для рядов, описывающих прибытие вагонов с нефтью и нефтепродуктами.

Из графиков следует, что средняя ошибка прогноза не падает с увеличением длины предыстории. Поэтому в качестве критерия для выбора этого параметра была использована стабилизация распределения, описываемого гистограммой. Для этого вычислялось расстояние Кульбака-Лейблера между парой распределений, построенных по наборам точек, отличающихся на 10 точек:

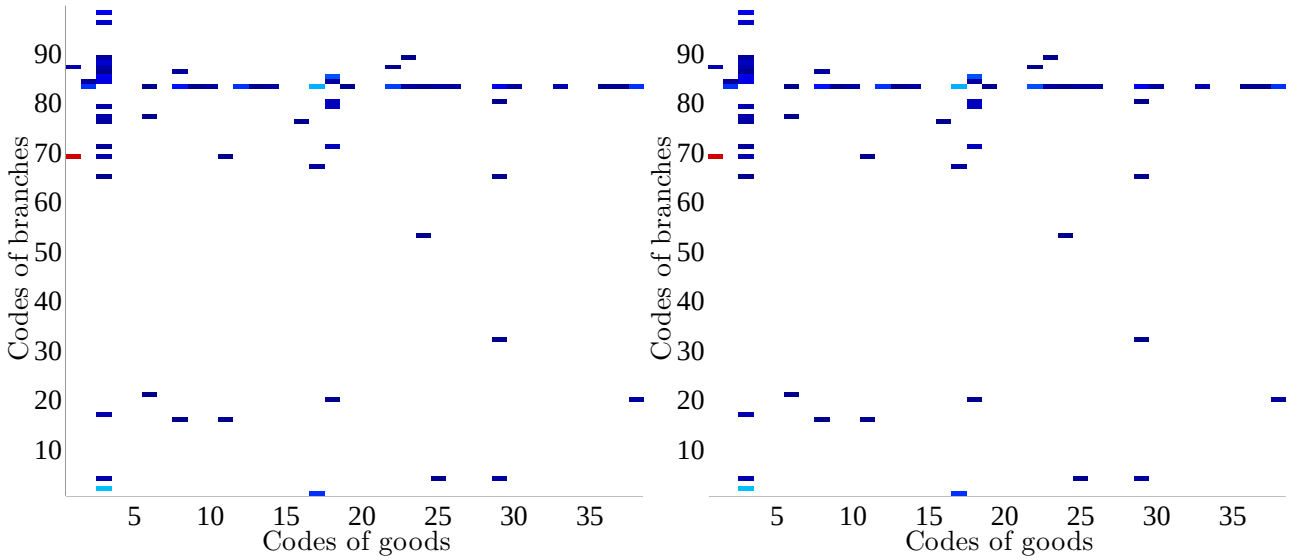
$$dist(p_1, p_2) = \sum_{i=1}^{k} p_1(i) \ln \left(\frac{p_1(i)}{p_2(i)} \right),$$

где p_1, p_2 — плотности дискретных распределений, задаваемых гистограммами, i — принимаемые случайной величиной значения.



(a) Прибытие, неделя

(b) Прибытие, месяц

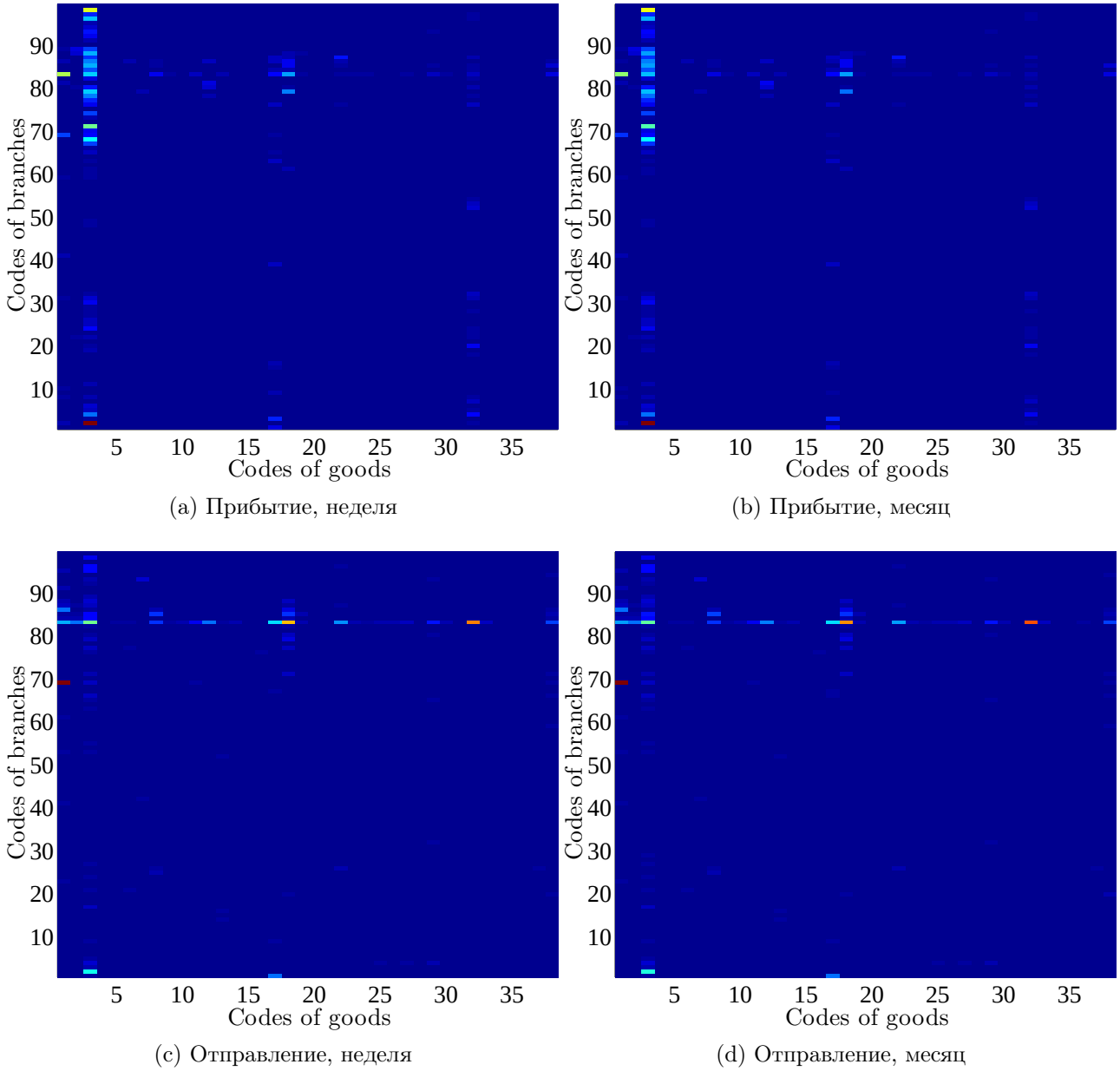


(c) Отправление, неделя

(d) Отправление, месяц

Рис. 10: Средняя ошибка при прогнозировании ARMA

На рис. 8 изображен график зависимости расстояния Кульбака-Лейблера между парой гистограмм от длины предыстории. По горизонтали отложено число точек, использованных при построении гистограммы, по вертикали — расстояние между двумя последовательно построенными гисторгаммами. Точки последовательно набираются, начиная с конца временного ряда. Границы интервалов для гистограмм фиксируются по границам интервалов оптимальной гистограммы, построенной по всему временному ряду. Для дальнйших экспериментов была выбрана длина предыстории $H = 120$ точек, так как такого количества достаточно, как это следует из графиков, для получения расстояния между двумя последовательно построенными гистограммами не более 0.05.



(a) Прибытие, неделя

(b) Прибытие, месяц

(c) Отправление, неделя

(d) Отправление, месяц

Рис. 11: Средняя ошибка при непараметрическом прогнозировании с абсолютной функцией потерь

При построении прогнозов были использованы свертки с тремя функциями потерь:

1) $L(z,x) = (z-x)^2$;
2) $L(z,x) = |z-x|$;
3) $L(z,x) = \begin{cases} 0, & \text{если } |z-x| < a; \\ |z-x| - a, & \text{если } |z-a| \geqslant a, \text{ где } a = 19. \end{cases}$

Значение параметра a соответствует разности числа вагонов в соседних столбцах гистограммы.

На рис. 9 изображены свертки гистограммы с различными функциями потерь. По горизонтали отмечено количество вагонов, соответствующее среднему значению интервала гистограммы, по вертикали — количество значений ряда, попавших в интервал, и значения сверток гистограммы с функциями потерь. Столбцы, соответствующие значениям суммы квадратов, уменьшены в 10^4 раз, соответствующие сумме модулей и трапеции — в 10^2 раз.

Сравнение непараметрического алгоритма прогнозирования с алгоритмом ARMA. При выполнении вычислительного эксперимента были вычислены средние ошибки (7) прогнозирования непараметрического алгоритма для прогноза на день, неделю и месяц по каждой ветке и каждой категории груза. Результаты сравнивались со средней ошибкой, получаемых с помощью модели ARMA для рядов, в которых ненулевых значений не менее $\frac{1}{5}$ от числа всех значений ряда. При меньшем количестве ненулевых значений модель ARMA работает некорректно. Во избежании деления на ноль при вычислении средней ошибки ко всем элементам рядов было прибавлено число 100.

На рис. 10 показаны средние ошибки, даваемые при прогнозе алгоритмом ARMA для рядов, в которых не менее $\frac{1}{5}$ ненулевых значений. Ряды, для которых прогноз не строился, отмечены белым цветом. Значение ошибки показано цветом ячейки: чем больше значение, тем более красный оттенок. По горизонтали отложены коды грузов, по вертикали — коды веток.

На рис. 11 показаны средние ошибки, даваемые при прогнозе алгоритмом непараметрического прогнозирования с абсолютной функцией потерь. Значение ошибки показано цветом ячейки: чем больше значение, тем более красный оттенок. Оси соответствуют тем же величинам, что и на предыдущем рисунке.

Табл. 12 содержит средние ошибки прогнозирования в процентах для модели ARMA и непараметрической модели с тремя рассматриваемыми функциями потерь для рядов, описывающих прибытие вагонов с различными типами грузов на 83 ветку. Табл. 13 содержит аналогичную информацию для отправления вагонов с 83 ветки. Первый столбец обеих таблиц содержит коды грузов, второй — информацию о стационарности рядов. Если в ячейке стоит 1, то ряд нестационарный, если 0, то стационарный. Из анализа результатов, представленных в таблицах, можно сделать вывод, что применение непараметрической прогностической модели для прогнозирования нестационарных временных рядов возможно, но обеспечивает меньшую точность прогноза, чем при прогнозировании стационарных рядов.

Заключение

Предложен алгоритм непараметрического прогнозирования загруженности железнодорожных узлов РЖД, основанный на свертке эмпирической плотности распределения значений временного ряда с функцией потерь. Проведен анализ структуры входных данных, на основе которого выбраны параметры модели. Проведен сравнительный анализ результатов предложенного алгоритма и алгоритма ARMA. Основным преимуществом

Code	N/S	Week				Month			
		ARMA	hist(SSE)	hist(abs)	hist(trap)	ARMA	hist(SSE)	hist(abs)	hist(trap)
0	0	NaN	0,05	0,05	0,05	NaN	0,05	0,05	0,05
1	1	NaN	34,63	33,23	33,46	NaN	34,05	33,02	33,28
2	0	NaN	0,33	0,29	0,29	NaN	0,34	0,30	0,30
3	1	NaN	20,88	19,87	20,02	NaN	20,75	20,11	19,89
4	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
6	0	NaN	0,52	0,38	0,38	NaN	0,53	0,38	0,38
7	0	NaN	0,58	0,38	0,38	NaN	0,58	0,38	0,38
8	0	NaN	0,63	0,42	0,42	NaN	0,64	0,41	0,41
9	1	12,92	8,27	5,88	12,76	15,43	8,94	6,04	13,50
10	0	2,04	1,23	1,01	6,62	2,26	1,26	1,02	7,07
11	0	NaN	0,40	0,26	0,26	NaN	0,40	0,26	0,26
12	0	NaN	3,25	3,51	6,88	NaN	3,26	3,57	7,04
13	0	NaN	0,87	0,79	0,79	NaN	0,88	0,79	0,79
14	0	4,33	2,43	2,31	2,64	4,80	2,42	2,34	2,43
15	0	1,10	0,69	0,49	0,49	1,22	0,71	0,51	0,51
16	0	NaN	0,20	0,15	0,15	NaN	0,20	0,15	0,15
17	0	NaN	0,54	0,41	0,41	NaN	0,52	0,39	0,39
18	0	15,00	8,33	7,47	10,19	16,22	8,45	7,56	10,23
19	1	30,10	19,45	16,89	17,77	34,86	20,53	17,77	18,71
20	0	2,12	1,42	1,17	7,01	2,37	1,43	1,17	7,44
21	0	NaN	0,19	0,12	0,12	NaN	0,20	0,12	0,12
22	0	NaN	0,44	0,33	0,33	NaN	0,45	0,34	0,34
23	0	2,85	1,66	1,33	1,33	2,87	1,67	1,33	1,30
24	0	2,24	1,22	1,11	1,21	2,39	1,21	1,11	1,21
25	0	2,33	1,35	1,18	1,26	2,60	1,36	1,18	1,28
26	0	1,61	0,97	0,65	0,65	1,79	0,98	0,66	0,66
27	0	1,25	0,78	0,53	0,53	1,39	0,79	0,55	0,55
28	0	2,74	1,56	1,09	1,09	2,83	1,56	1,08	1,08
29	0	NaN	0,32	0,23	0,23	NaN	0,33	0,23	0,23
30	0	6,66	3,90	3,57	9,54	7,48	3,99	3,63	9,77
31	0	2,92	1,56	1,51	1,99	3,19	1,55	1,52	2,04
33	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
34	0	6,38	3,60	3,01	3,92	6,96	3,61	2,90	4,00
35	0	1,85	0,94	0,89	0,87	1,95	0,92	0,88	0,86
36	0	0,99	0,67	0,47	0,47	1,04	0,67	0,47	0,47
38	0	1,77	1,07	0,94	0,95	1,83	1,04	0,93	0,93
39	0	1,25	0,71	0,55	0,55	1,34	0,72	0,56	0,56
42	0	NaN	0,55	0,51	0,51	NaN	0,52	0,50	0,50
43	0	NaN	5,63	5,16	10,09	NaN	5,93	5,30	10,45

Рис. 12: Средний процент ошибки при прогнозе прибытия

Code	N/S	Week				Month			
		ARMA	hist(SSE)	hist(abs)	hist(trap)	ARMA	hist(SSE)	hist(abs)	hist(trap)
0	0	NaN	0,05	0,05	0,05	NaN	0,05	0,05	0,05
1	1	NaN	34,63	33,23	33,46	NaN	34,05	33,02	33,28
2	0	NaN	0,33	0,29	0,29	NaN	0,34	0,30	0,30
3	1	NaN	20,88	19,87	20,02	NaN	20,75	20,11	19,89
4	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
6	0	NaN	0,52	0,38	0,38	NaN	0,53	0,38	0,38
7	0	NaN	0,58	0,38	0,38	NaN	0,58	0,38	0,38
8	0	NaN	0,63	0,42	0,42	NaN	0,64	0,41	0,41
9	1	12,92	8,27	5,88	12,76	15,43	8,94	6,04	13,50
10	0	2,04	1,23	1,01	6,62	2,26	1,26	1,02	7,07
11	0	NaN	0,40	0,26	0,26	NaN	0,40	0,26	0,26
12	0	NaN	3,25	3,51	6,88	NaN	3,26	3,57	7,04
13	0	NaN	0,87	0,79	0,79	NaN	0,88	0,79	0,79
14	0	4,33	2,43	2,31	2,64	4,80	2,42	2,34	2,43
15	0	1,10	0,69	0,49	0,49	1,22	0,71	0,51	0,51
16	0	NaN	0,20	0,15	0,15	NaN	0,20	0,15	0,15
17	0	NaN	0,54	0,41	0,41	NaN	0,52	0,39	0,39
18	0	15,00	8,33	7,47	10,19	16,22	8,45	7,56	10,23
19	1	30,10	19,45	16,89	17,77	34,86	20,53	17,77	18,71
20	0	2,12	1,42	1,17	7,01	2,37	1,43	1,17	7,44
21	0	NaN	0,19	0,12	0,12	NaN	0,20	0,12	0,12
22	0	NaN	0,44	0,33	0,33	NaN	0,45	0,34	0,34
23	0	2,85	1,66	1,33	1,33	2,87	1,67	1,33	1,30
24	0	2,24	1,22	1,11	1,21	2,39	1,21	1,11	1,21
25	0	2,33	1,35	1,18	1,26	2,60	1,36	1,18	1,28
26	0	1,61	0,97	0,65	0,65	1,79	0,98	0,66	0,66
27	0	1,25	0,78	0,53	0,53	1,39	0,79	0,55	0,55
28	0	2,74	1,56	1,09	1,09	2,83	1,56	1,08	1,08
29	0	NaN	0,32	0,23	0,23	NaN	0,33	0,23	0,23
30	0	6,66	3,90	3,57	9,54	7,48	3,99	3,63	9,77
31	0	2,92	1,56	1,51	1,99	3,19	1,55	1,52	2,04
33	0	NaN	0,03	0,03	0,03	NaN	0,03	0,03	0,03
34	0	6,38	3,60	3,01	3,92	6,96	3,61	2,90	4,00
35	0	1,85	0,94	0,89	0,87	1,95	0,92	0,88	0,86
36	0	0,99	0,67	0,47	0,47	1,04	0,67	0,47	0,47
38	0	1,77	1,07	0,94	0,95	1,83	1,04	0,93	0,93
39	0	1,25	0,71	0,55	0,55	1,34	0,72	0,56	0,56
42	0	NaN	0,55	0,51	0,51	NaN	0,52	0,50	0,50
43	0	NaN	5,63	5,16	10,09	NaN	5,93	5,30	10,45

Рис. 13: Средний процент ошибки при прогнозе отправления

непараметрического алгоритма по сравнению с алгоритмом ARMA является его применимость для прогнозирования стационарных временных рядов с большим количеством одинаковых значений, в том числе и нулевых. Проверена практическая применимость предложенного алгоритма для прогнозированияя нестационарных временных рядов.

Литература

[1] Koenker Jr., Bassett G. Regression Quantiles//Econometrica. 1978. Vol. 46. № 1. P. 33–50.

[2] Постникова Е. Квантильная регрессия. — Новосибирск: НГУ, 2006.

[3] Хардле В. Прикладная непараметрическая регрессия. — М: Мир, 1993.

[4] Шурыгин А. М. Прикладная стохастика: робастность, оценивание, прогноз. — М: Финансы и статистика, 2000.

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

[6] Магнус Я. Р., Катышев П. К., Персецкий А. А. Эконометрика: начальный курс. — М: Дело, 2004.

[7] Thiesing F. M., Vornberger O. Sales Using Neural Networks//Lecture Notes in Computer Science. 1997. Vol. 1226. P. 321–328.

[8] Gheyas I. A., Smith L. S. Neural network approach to time series Forecasting//Proceedings of the World Congress on Engineering. 2009. Vol. 2. P. 245–253.

[9] Cortez P., Rotcha M., Neves J. Evolving time series forecasting ARMA models//Journal of Heuristics. 2004. Vol. 10(4). P. 419–429.

[10] Nochai R., Nochai T. ARIMA model for forecasting oil palm price. Proceedings of the 2nd IMT-GT Regional conference of mathematics, statistics and applications. University Sains Malaisia, Penang. June 13–15, 2006.

[11] Shumway R. H., Stoffer D. S. Time series analysis and its applications with R Examples. Springer. 2006.

[12] Scott D. W. On optimal and data-based histograms. Biometrika, 1979. Vol. 66(3). P. 605—610.

Комбинированный порождающий и разделяющий подход в задачах классификации с малой выборкой*

Животовский Н. К.
nikita.zhivotovskiy@phystech.edu
Московский физико-технический институт

В работе рассмотрены два статистических подхода к решению задачи классификации и способ их комбинации, предназначенный для оценки параметров классификатора по выборкам различной мощности. Для случая, когда объекты в классах имеют многомерное нормальное распределение, построена комбинированная модель, сочетающая в себе порождающий и разделяющие подходы к задачам классификации. В серии экспериментов показано, что при ограничениях на длину обучающей выборки использование этой модели может приводить к уменьшению вероятности ошибки получаемого классификатора по сравнению с чисто порождающими или разделяющими моделями.

Ключевые слова: *классификация, порождающая и разделяющая модели, логистическая регрессия.*

Combined generative and descriminative approach for calssification with a small learning set*

Zhivotovskiy N. K.
Moscow Institute of Physics and Tecnology

This paper deals with two statistical approaches to solving classification problems and way of their combination designed to evaluate the parameters of a classifier for samples of different cardinality. The combined descriminative and generative model was built for the case of the multivariate normal distribution of objects within classes. This model shows lower probability of error of classificator as compared with one obtained purely from generative or descriminative model when restrictions are put on the size of the learning set.

Keywords: *classification, generative and descriminative approaches, logistic regression.*

Введение

Исследуется комбинированный порождающий и разделяющий подход в задачах классификации, описанный в [1]. Задача классификации заключается в нахождении оптимального, например, с точки зрения вероятности ошибки зрения правила, которое относит объекты, представленные точками конечномерного действительного векторного пространства к одному из конечного числа классов. Построенное правило называется *классификатором*. Классификатор выбирается из множества правил, которые называются *моделью*. Модель описывается в виде параметрически заданного семейства функций, отображающих множество объектов во множество классов. При оценке параметров модели с целью выбора оптимального классификатора используется заранее известная конечная выборка объектов, называемая *обучающей*, для которой уже известен класс каждого из входящих в нее объектов.

Научный руководитель В.В. Стрижов
Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00709.

Используемые модели подразделяется на *разделяющие* и *порождающие* [1, 2]. Оценка параметров в разделяющих моделях можно рассматривать как подбор таких значений параметров модели, которые максимизирует правдоподобие обучающей выборки по отношению к вероятности класса [2]. Классификатор в таком случае относит объект к его наиболее вероятному классу. Альтернативный подход, называющийся *порождающим* [1, 2], заключается в максимизации правдоподобия совместного распределения объектов и классов, а затем в использовании формулы Байеса для нахождения вероятности отношения объекта к классу.

В работе [3] производится сравнение обоих подходов на примере логистической регрессии [2], для которой параметры оцениваются исходя из разделяющего подхода, и наивного Байесовского классификатора [2], для которого оценка параметров максимизирует функцию совместного правдоподобия объектов и классов. Результаты теоретического и экспериментального исследований подтверждают, что с ростом длины обучающей выборки разделяющий подход приводит к меньшей вероятности ошибки, т. е. к лучшему качеству классификации.

Однако, во-первых, для получения меньшей вероятности ошибки разделяющий подход требует большую длину обучающей выборки, в то время как порождающий подход достигает своего асимптотического по длине обучающей выборки минимума вероятности ошибки гораздо быстрее. Во-вторых, для малых длин выборок на 14 из 15 рассмотренных в статье задачах из репозитория UCI порождающий подход дает в среднем лучшее качество классификации.

Таким образом, ни разделяющий, ни порождающий подход не является строго предпочтительным для всех длин выборок и для всех задач. Поэтому в целях улучшения качества классификации в ряде работ исследуется идея комбинированного подхода. В [1] и [4] для модельных данных, а также для задач классификации изображений, показано, что с помощью выбора подходящей модели комбинация двух подходов может улучшать качество классификации по сравнению с каждым из подходов по отдельности. В частности, в [1] комбинированный подход к оценке параметров модели заключается в замене правдоподобия обучающей выборки на выпуклую комбинацию логарифмов правдоподобий, относящихся соответственно к разделяющему и порождающему подходам. Альтернативные подходы к построению комбинированных моделей для задач классификации изображений приводятся, например, в [5].

Постановка задачи

Пусть $\{\mathbf{x}_i, y_i\}_{i=1}^{\ell}$ — конечная обучающая выборка, выбранная независимо из некоторого неизвестного совместного распределения объектов и классов, а $\{1, -1\}$ — множество классов. Будем считать, что $P(y|\mathbf{x}, \boldsymbol{\theta})$ — вероятность принадлежности объекта $\mathbf{x}$ классу y и $p(y, \mathbf{x}|\boldsymbol{\theta})$ — совместная плотность распределения объектов и классов задаются общим набором параметров $\boldsymbol{\theta}$, априорно неизвестных.

Согласно [1] предполагается, что искомые значения параметров максимизируют выпуклую оболочку логарифма разделяющего правдоподобия обучающей выборки

$$L_D = \sum_{i=1}^{\ell} \log\bigl(p(y_i|\mathbf{x}_i, \boldsymbol{\theta})\bigr) \tag{1}$$

зависящего от вероятности принадлежности объекта выборки к классу и логарифма ее порождающего правдоподобия

$$L_G = \sum_{i=1}^{\ell} \log\bigl(p(y_i, \mathbf{x}_i|\boldsymbol{\theta})\bigr) \tag{2}$$

зависящего от совместной плотности объектов и классов. Общая формула имеет следующий вид:

$$\lambda L_D + (1-\lambda)L_G = \lambda \sum_{i=1}^{\ell} \log\bigl(p(y_i|\mathbf{x}_i, \boldsymbol{\theta})\bigr) + (1-\lambda) \sum_{i=1}^{\ell} \log\bigl(p(y_i, \mathbf{x}_i|\boldsymbol{\theta})\bigr), \quad \lambda \in [0,1] \tag{3}$$

Тогда поиск оптимальных значений параметров заключается в максимизации этой взвешенной суммы правдоподобий:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} (\lambda L_G + (1-\lambda)L_D)$$

Выбор параметра λ определяет значимость каждого из подходов.

Построение комбинации правдоподобий

В качестве основной модели, с помощью которой будет иллюстрироваться комбинированный подход. в данной работе принята логистическая регрессия [2]. В этой модели предполагается, что вероятность принадлежности объекта $\mathbf{x}$ к классу 1 задается в виде формулы:

$$P(1|\mathbf{x}) = \sigma((\mathbf{w}, \mathbf{x})),$$

где $\sigma(z) = \frac{1}{1+\exp(-z)}$ — сигмоидная функция, определенная для всех действительных z. Оценка параметров в случае логистической регрессии заключается в поиске такого значения вектора параметров $\mathbf{w}$, которое максимизирует логарифм правдоподобия обучающей выборки:

$$L = \sum_{i=1}^{\ell} \log\bigl(y_i \sigma(\mathbf{w}, \mathbf{x}_i)\bigr) \to \max_{\mathbf{w}}$$

Оценка параметров этой модели соответствует разделяющему подходу (1). В данной работе чисто разделяющая функция правдоподобия (1) будет заменена на выпуклую комбинацию разделяющего и порождающего (3).

Предполагается, что объекты $\mathbf{x} \in \mathbb{R}^n$, а функции правдоподобия $p_y(\mathbf{x})$ (плотности распределения объектов при фиксированном классе обозначаются соответственно $p_1(\mathbf{x})$ и $p_{-1}(\mathbf{x})$) имеют многомерное нормальное распределение со средними $\boldsymbol{\mu}_y$ и ковариационной матрицей $\boldsymbol{\Sigma}$, общей для обоих классов. Пусть P_1 — априорная вероятность класса $+1$, тогда вероятность класса -1 равна $1 - P_1$.

В таком случае можно рассчитать апостериорные вероятности классов. Заметим, что функция правдоподобия классов в нашем случае имеет вид:

$$p_y(\mathbf{x}) = \exp\bigl(\boldsymbol{\mu}_y^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_y^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y - \tfrac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{n}{2}\ln(2\pi) - \tfrac{1}{2}\ln(|\boldsymbol{\Sigma}|)\bigr).$$

Таким образом,

$$\frac{P(+1|\mathbf{x})}{P(-1|\mathbf{x})} = \frac{P_1 p_1(\mathbf{x})}{(1-P_1)p_{-1}(\mathbf{x})} = \frac{P_1}{(1-P_1)} \exp\bigl((\mathbf{w}, \mathbf{x}) + c\bigr),$$

где $\mathbf{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T \boldsymbol{\Sigma}^{-1},\ c = (\mathbf{w}, \boldsymbol{\mu}_1 + \boldsymbol{\mu}_{-1})$. Так как классов всего два, то

$$P(+1|\mathbf{x}) + P(-1|\mathbf{x}) = 1$$

Отсюда, с учетом полученного равенства, получаем:

$$\mathsf{P}(1|\mathbf{x}) = \frac{1}{1 + \frac{1-P_1}{P_1}\exp(-(\mathbf{w},\mathbf{x}) - c)}, \quad \mathsf{P}(-1|\mathbf{x}) = \frac{1}{1 + \frac{P_1}{1-P_1}\exp((\mathbf{w},\mathbf{x}) + c)}.$$

Формулой Байеса позволяется получить совместную плотность распределения $p(y, \mathbf{x})$:

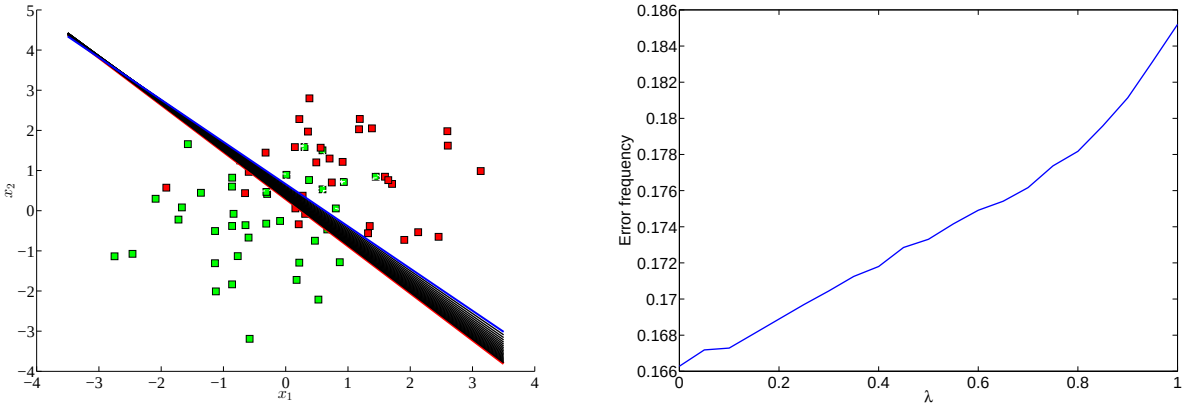
$$p(y, \mathbf{x}) = p(y|\mathbf{x})\big(P_1 p_1(\mathbf{x}) + (1 - P_1)p_{-1}(\mathbf{x})\big).$$

Подстановка полученных результатов в формулу для выпуклой комбинации правдоподобий дает:

$$L_\lambda = -\sum_{i=1}^{\ell}\Big(\log\big(1 + (\tfrac{1-P_1}{P_1})^{y_i}\exp(-y_i((\mathbf{w},\mathbf{x}_i) + c))\big) + (1-\lambda)\log\big(P_1 p_1(\mathbf{x}_i) + (1 - P_1)p_{-1}(\mathbf{x}_i)\big)\Big).$$

Стоит отметить, что L_λ зависит лишь от $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}, P_1$. Таким образом, получена оптимизационная задача:

$$L_\lambda \ \to \ \max_{\boldsymbol{\mu_1},\boldsymbol{\mu_{-1}},\boldsymbol{\Sigma},P_1}.$$



(a) Обучающая выборка и разделяющие прямые, соответствующие разным значениям λ

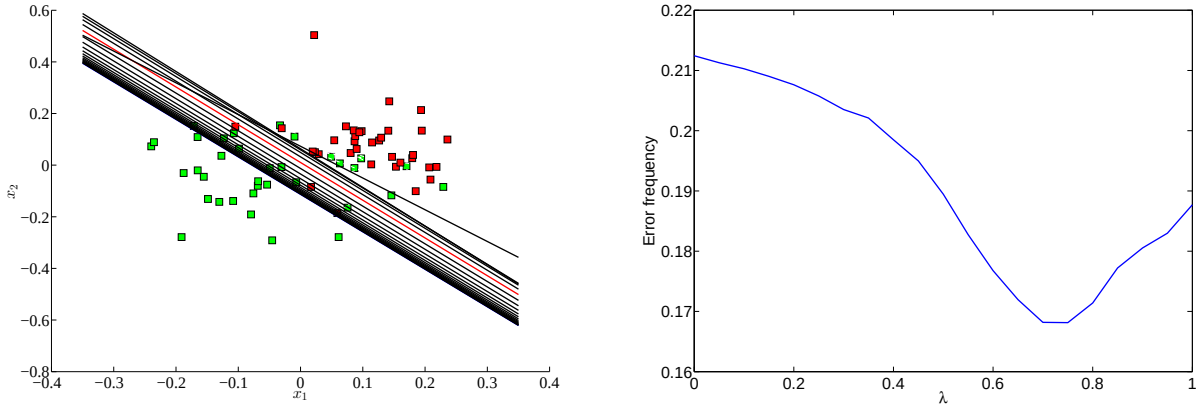
(b) Частота ошибок классификатора в зависимости от значения параметра λ

Рис. 1: Случай разреженных классов

Вычислительный эксперимент

Для иллюстрации комбинированного подхода к обучению производится серия экспериментов. Для двух классов генерируются обучающая выборка, выбранная из двумерного мерного нормального распределения. Ради упрощения вычислений в эксперименте предполагается, что ковариационная матрица $\boldsymbol{\Sigma}$ известна и является диагональной с σ^2 на диагонали. Параметр σ^2 будет изменяться в эксперименте и позволит задавать дисперсии объектов в классах. Предполагается, что априорные вероятности классов равны. Координаты средних значений правдоподобий классов $\boldsymbol{\mu}_y$ при этом равны $(-0.7\sigma, -0.7\sigma)$ для

класса $y = -1$ и $(0.7\sigma, 0.7\sigma)$ для класса $y = 1$. Выбор таких средних позволяет с одной стороны достичь некоторого смешения классов, а с другой стороны производит их кластеризацию вокруг удаленных друг от друга средних. При этом чем меньше дисперсия σ^2, тем меньше и расстояние между классами, т. е. выборка не будет линейно разделима практически при всех значениях σ^2. Множитель 0.7 задает соотношение между дисперсией классов и их средним и характеризует степень смешения классов. Чем меньше его значение, тем меньше расстояние между классами и тем хуже они разделяются прямой.



(a) Обучающая выборка и разделяющие прямые, соответствующие разным значениям λ

(b) Частота ошибок классификатора в зависимости от значения параметра λ

Рис. 2: Случай классов с малой дисперсией

Тем не менее параметры $\boldsymbol{\mu}_y$ считаются неизвестными и оцениваются при максимизации правдоподобия. Для эксперимента создавалась генеральная выборка из $N = 10000$ прецедентов из описанного распределения. Из нее случайным образом выбиралась обучающая выборка из ℓ элементов. Пусть $\mathbf{x}_i$ – i-й объект выборки, а выбранные ℓ объектов имеют соответственно номера $1, \ldots, \ell$. Индикатор ошибки классификатора на $\mathbf{x}_i$ обозначается как $I(\mathbf{x}_i)$. Тогда критерием качества классификатора будет минимум частоты ошибок на объектах, не попавших в обучающую выборку:

$$\frac{\sum\limits_{i=\ell+1}^{N} I(\mathbf{x}_i)}{N - \ell} \to \min$$

Аналогично результату из [3] на больших обучающих выборках разделяющий подход, соответствующий $\lambda = 0$, показывает лучшее качество классификации, т. е. меньшую частоту ошибок на оставшихся $N - \ell$ объектах.

Выборки, длина которых меньше $30 - 40$ объектов, не позволяют понять важность каждого из подходов, так как дисперсия частоты ошибок получаемого классификатора слишком велика. Поэтому в качестве промежуточного значения выбрано $\ell = 70$.

Таким образом, $L_\lambda = L_\lambda(\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1})$. Для поиска параметров распределения, максимизирующих введенное правдоподобие, был использован оптимизационный toolbox yalmip языка Matlab. В случае комбинированного правдоподобия, как и в случае простой логистической регрессии, градиентные методы успешно находят локальные максимумы.

Случай разреженных классов.

Сначала рассматривается случай $\sigma = 1$.

На рис. 81 изображено множество разделяющих прямых, которые получаются при различных значениях параметра λ. По осям отложены координаты объектов. Можно заметить, что в данном случае происходит движение разделяющей прямой при изменении λ. Верхняя прямая в пучке при этом соответствует разделяющему подходу, а нижняя — порождающему. Красным цветом выделена прямая, которая доставляет наименьшую частоту ошибок на генеральной выборке. Синим цветом — доставляющая наибольшую частоту ошибок.

Частота ошибок на генеральной выборке, отложенная по оси ординат, в зависимости от λ изображена на рис. 82.

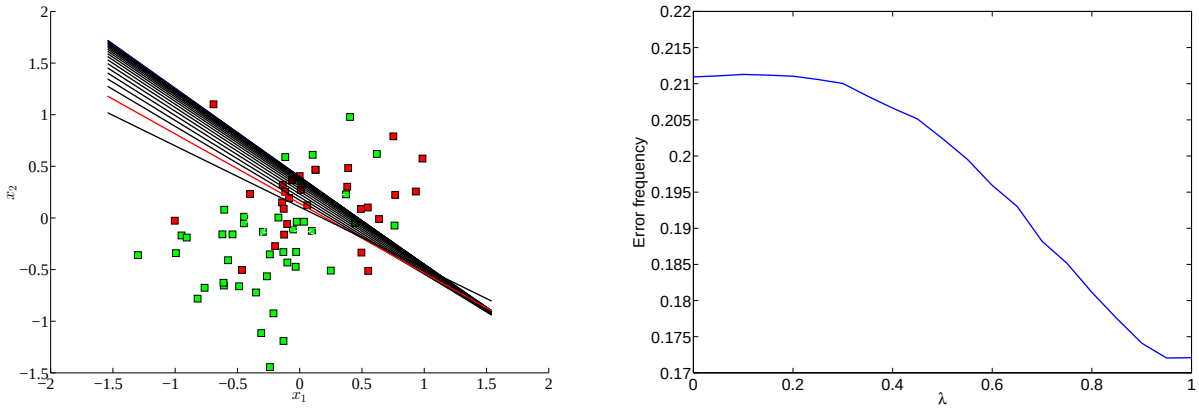
В случае, когда дисперсия классов σ^2 велика, большинство объектов удалено от всего множества разделяющих прямых. Таким образом, использование взвешенной функции правдоподобия практически не изменяет частоту ошибок получаемого классификатора.

Случай классов с малой дисперсией.

Пусть теперь $\sigma = 0.1$. Как показывают рис. 83 и рис. 84 в этом случае комбинированный подход позволяет существенно улучшить качество классификации.

Однако дальнейшее уменьшение параметра σ не позволяет получить уменьшение частоты ошибки при использовании комбинированного подхода. Более того, прямые, соответствующие значениям λ не равным нулю или единице, даже ухудшают качество классификации.

Случай средней разреженности. В качестве промежуточного случая рассматривается $\sigma = 0.44$. Соответствующие этому случаю иллюстрации изображены на рисунках 85 и 86.



(a) Обучающая выборка и разделяющие прямые, соответствующие разным значениям λ

(b) Частота ошибок классификатора в зависимости от значения параметра λ.

Рис. 3: Случай средней разреженности

В этом случае комбинированный подход позволяет получить некоторое улучшение качества классификации.

Заключение

Серия экспериментов показывает, что для малых обучающих выборок комбинированный подход к оценке параметров позволяет в некоторых случаях улучшить качество клас-

сификации. В отличие от логистической регрессии результат классификации при комбинированном подходе существенно зависит от масштаба вектора. Действительно, выбор параметра дисперсии просто изменяет координаты объекта, являющегося действительным вектором, в одно и то же число раз.

В случае больших значений дисперсий малая часть объектов попадает в окрестность тех прямых, которые разделяют классы при разных значениях параметра, регулирующего вклад каждого из подходов. Поэтому качество классификации практически не меняется, если использовать комбинированный подход.

В то же время при меньших значениях дисперсии удается существенно улучшить качество классификации, используя комбинированный подход.

Литература

[1] *Bishop C. M. and Lasserre J.* Generative or discriminative? Getting the best of both worlds (2007) // Bayesian Statistics 8. С. 3–24.

[2] *Bishop C. M.* Pattern recognition and machine learning (2006) // Springer, Series: Information Science and Statistics 8. — С. 740 p.

[3] *Ng, A. Y. and Jordan, M. I.* On discriminative vs. generative: A comparison of logistic regression and naive Bayes (2002) // Advances in Neural Information Processing Systems 14. Cambridge, MA: The MIT Press. С. 841–848.

[4] *Lasserre J., Bishop C. M., and Minka T.* Principled hybrids of generative and discriminative models (2006) // Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 1. С. 87–94.

[5] *Perina A., Cristani M., Castellani U., Murino V., and Jojic N.* A hybrid generative/discriminative classification framework based on free-energy terms (2009) // Computer Vision, 2009 IEEE 12th International Conference. С. 2058–2065 .

Алгоритм выделения устойчивых отражателей на спутниковых радиолокационных снимках земной поверхности*

Василейский А. С.[1], ***Карацуба Е. А.***[2], ***Карелов А. И.***[1], ***Кузнецов М. П.***[3], ***Рейер И. А.***[2]

A.Vasileisky@gismps.ru, karacuba@mi.ras.ru, A.Karelov@gismps.ru, mikhail.kuznecov@phystech.edu, reyer@forecsys.ru

1 — Центр внедрения космических технологий ОАО «НИИАС»

2 — Вычислительный центр им. Дородницына РАН

3 — Московский физико-технический институт

Исследуется проблема выделения устойчивых отражателей радиолокационного сигнала, расположенных на поверхности земли. Устойчивые отражатели определяются по данным радиолокационных спутниковых снимков, содержащих амплитудную и фазовую составляющие. Определение координат отражателей происходит по амплитудной составляющей. Фазовая составляющая используется для определения движения отражателей с течением времени вследствие сдвига земной поверхности. Предложен алгоритм выделения отражателей как ярких пятен на амплитудной составляющей. Алгоритм проиллюстрирован синтетическими и реальными данными. В работе описан метод обработки спутниковых снимков, построения и проверки адекватности синтетических данных и процедура построения системы устойчивых отражателей.

Ключевые слова: *радиолокация, синтезированная апертура, SAR-интерферометрия, устойчивые отражатели, LoG-детектор.*

The algorithm of persistent scatterers detection on the satellite radar images of the earth surface*

Vasileisky A. S.[1], ***Karatsuba E. A.***[2], ***Karelov A. I.***[1], ***Kuznetsov M. P.***[3], ***Reyer I. A.***[2]

1 — Space Technology Application Center of NIIAS

2 — Dorodnitsyn Computing Centre

3 — Moscow Institute of Physics and Technology

We consider a problem of the radar signal persistent scatterers detection on the earth surface. To detect the scatterers we use satellite SAR images consisting of the amplitude and phase components. To identify scatterers coordinates we use amplitude component. Phase component is used to determine scatterers movement due to the terrain shifts. We propose a blob detection algorithm to find the scatterers. To illustrate the algorithm we use synthetic and real data. We describe a method of the satellite images processing, a method of the synthetic data construction and verification and method of the persistent scatterers system detection.

Keywords: *radiolocation, synthetic aperture radar, SAR interferometry, persistent scatterers, LoG-detector.*

Введение

Мониторинг состояния объектов железнодорожной инфраструктуры включает контроль стабильности пути, зданий и сооружений с использованием различных методов из-

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-13160.

мерений. Раннее обнаружение отклонений геометрических параметров пути от заданных значений, деформаций и потери устойчивости искусственных сооружений, связанных как с эксплуатационными нагрузками, так и с воздействием экзогенных процессов, необходимо для поддержания исправного состояния железнодорожной инфраструктуры и обеспечения безопасности перевозок. Дистанционные методы аэрокосмического мониторинга эффективно дополняют традиционный контроль состояния пути вагонами-путеизмерителями и контроль стабильности сооружений геодезическими методами за счет обнаружения потенциально опасных воздействий на прилегающих к объектам инфраструктуры территориях на оползневых участках, в зонах воздействия эрозионных процессов, при прохождении пути по территориям, подверженным карстовым процессам, и в зоне распространения многолетнемерзлых пород.

Данная работа посвящена описанию и анализу методов, позволяющих осуществлять контроль стабильности устойчивых отражателей радиолокационного сигнала, представляющих собой элементы наземных сооружений и открытые участки местности, по данным спутниковой радиолокационной съемки.

Для обнаружения изменений ландшафта, просадок и смещений земной поверхности, вызванных оползнями, землетрясениями, извержением вулканов и другими экзогенными геологическими процессами, применяется метод радиолокационной дифференциальной интерферометрии [1]. В его основе лежит съемка территории радарами с синтезированной апертурой (synthetic aperture radar - SAR) на борту космических аппаратов. SAR системы позволяют получать детальные радиолокационные изображения земной поверхности путем искусственного увеличения апертуры бортовой антенны.

Радиолокатор с синтезированной апертурой фиксирует амплитуду и фазу отраженного сигнала. Принцип интерферометрии заключается в восстановлении цифровой модели рельефа путем анализа фазовых компонент двух (или более) SAR-снимков одного и того же участка земной поверхности, сделанных с отстоящих точек одной орбиты. Дифференциальная интерферометрия применяется для получения информации об изменениях ландшафта.

Ранее в работах [1, 2, 3, 4, 5, 6] были исследованы методы обработки фазовых частей радиолокационных снимков для решения задачи интерферометрии. Сигнал, участвующий в формировании каждого пиксела, отражается от большого количества маленьких участков поверхности, которые могут обладать разными диэлектрическими свойствами. При этом фаза отраженного сигнала может быть абсолютно нескоррелирована для соседних пикселов, что затрудняет анализ фазовой картины и восстановление смещений поверхности [2].

Для решения этой проблемы было разработано понятие устойчивых отражателей [3, 4, 5, 6] радиолокационного сигнала. Устойчивыми отражателями является такой (разреженный) набор участков на поверхности, для которых амплитудно-фазовые характеристики сигнала при отражении незначительно меняются со временем. Такими участками могут являться, например, крыши зданий, сооружения или открытые участки грунта с неразвитой растительностью. Для последовательности снимков одной и той же поверхности выделяется набор устойчивых отражателей, и для каждого отражателя вычисляется изменение фазы отраженного сигнала. По этому изменению фаз делаются выводы об относительном сдвиге участков земной поверхности за время между съемками.

Для поиска на снимке устойчивых отражателей были предложены методы [6], основанные на построении вероятностной модели отраженного сигнала. Сигнал, отражающийся от элемента поверхности, представляется случайной величиной, для которой дисперсия

обратно пропорциональна показателю «устойчивости» отражателя. Для введенной таким образом вероятностной модели были предложены методы поиска отражателей, основывающиеся, например, на методе максимального правдоподобия.

В последующих работах были предложены методы поиска устойчивых отражателей на основе анализа графов [3], основывающиеся на переборе ребер, соединяющих все пары пикселов для двух изображений. При этом, исходя из минимизации некоторого функционала, среди всех ребер выбираются те, которые с наибольшей вероятностью соединяют пары устойчивых отражателей.

В данной работе предлагаются алгоритмы получения синтетических радиолокационных снимков с использованием измерений высот земной поверхности и выделения устойчивых отражателей по амплитудным компонентам снимков. Для выделения отражателей применяется LoG-детектор [7, 8]. Идея этого детектора заключается в последовательной свертке изображения с лапласианом гауссиан разного масштаба. По полученному набору сверток вычисляются максимумы функции отклика, соответствующие отражателям, а их размер определяется масштабом найденной гауссианы.

Проанализирована адекватность соответствия синтетических и реальных данных - радиолокационных снимков земной поверхности, полученных системой из четырех спутников COSMO-SkyMed, оснащенных SAR-аппаратурой. Изображения, полученные этой системой, используются для составления карт земной поверхности, контроля за береговыми линиями и предупреждения природных чрезвычайных ситуаций. Некоторые изображения находятся в открытом доступе: http://www.cosmo-skymed.it/en/index.htm.

Принцип радиолокационного синтеза апертуры

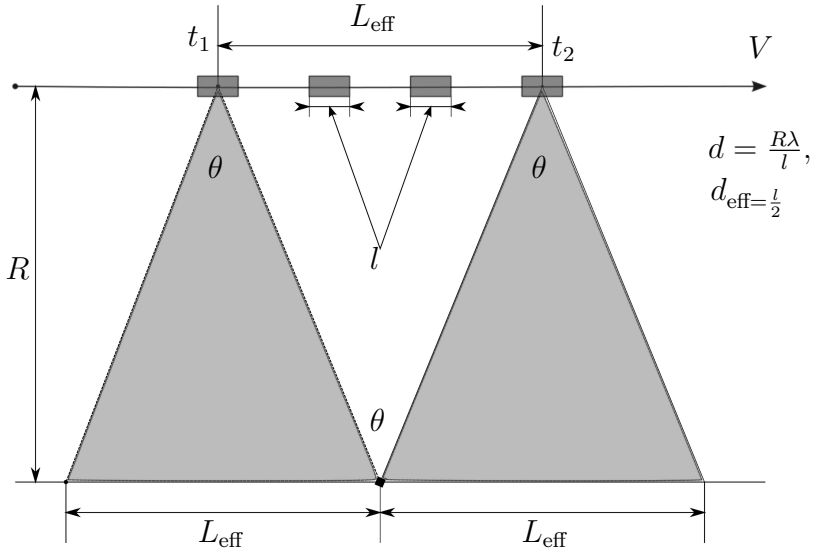


Рис. 1: Схема активной съемки земной поверхности с использованием космического радиолокатора с синтезированной апертурой

Основная идея принципа синтеза апертуры [1, 2] продемонстрирована на рис. 1. В общем случае, вследствие дифракции, разрешающая способность радиолокатора напрямую зависит от линейных размеров антенны:

$$\theta = \frac{\lambda}{l},$$

где l — размер антенны, λ — длина волны, θ — угол апертуры. Здесь и далее, величины λ, L, d, R измеряются в метрах, θ — в радианах. Детальность d радиолокационного изображения тем выше, чем больше размер антенны l и чем меньше расстояние до поверхности R:

$$d = \frac{R\lambda}{l}.$$

Таким образом, для получения высокой разрешающей способности (до нескольких сантиметров) необходимо располагать антеннами больших апертурных размеров, существенно превышающих реализуемые на реальных носителях. Для решения этой проблемы был разработан принцип искусственного синтеза апертуры при поступательном движении летательного аппарата [1, 2].

Этот принцип основан на перемещении излучателя вдоль требуемого апертурного размера L_{eff} со скоростью V при движении летательного аппарата, последовательно испускающего и принимающего отраженные от цели сигналы, а затем совместно обрабатывающего их. При этом появляются затраты времени на синтезирование, связанные с тем, что спутнику, летящему со скоростью V, необходимо пролететь расстояние L_{eff}, чтобы получить серию снимков, участвующих в формировании изображения, а также появляются вычислительные затраты на синтезирование снимков. Однако при этом достигается необходимый размер апертурного угла

$$\theta_{\text{eff}} = \frac{\lambda}{L_{\text{eff}}},$$

что позволяет добиться высокой разрешающей способности:

$$d_{\text{eff}} = \frac{l}{2},$$

где l — реальный апертурный размер антенны. На рис. 2 схематически изображен принцип эквивалентности реальной апертуры большого размера и синтезированной апертуры движущегося летательного аппарата. На рис. 2(а) показан радар с реальным размером апертуры L. На рис. 2(б) показан радар, искусственно синтезирующий апертуру размера L_{eff} в течение своего полета путем последовательного наложения снимков. Благодаря этому принципу в радаре на рис. 2(б) для получения разрешающей способности радара на рис. 2(в) требуется антенна гораздо меньшего размера.

Формат используемых данных.

Для иллюстрации работы предлагаемого алгоритма использованы SAR-снимки, полученные системой COSMO-SkyMed.

Данная система предусматривает SAR-съемку в трех различных режимах [9]:

Spotlight – режим съемки небольшого участка местности с высоким разрешением (антенна направлена на снимаемый участок в течение продолжительного времени);

Stripmap – режим съемки полосы местности со средним разрешением (фиксированное направление антенны);

ScanSAR – режим съемки большой территории (сканирование нескольких полос со сменой направления антенны) с низким разрешением.

Используемые изображения получены системой COSMO-SkyMed в подрежиме HIMAGE режима съемки Stripmap (съемка полосы с фиксированной конфигурацией радара, второй подрежим, PINGPONG, подразумевает съемку полосами с переключением поляризации). Ширина полосы съемки в подрежиме HIMAGE - приблизительно 40 км,

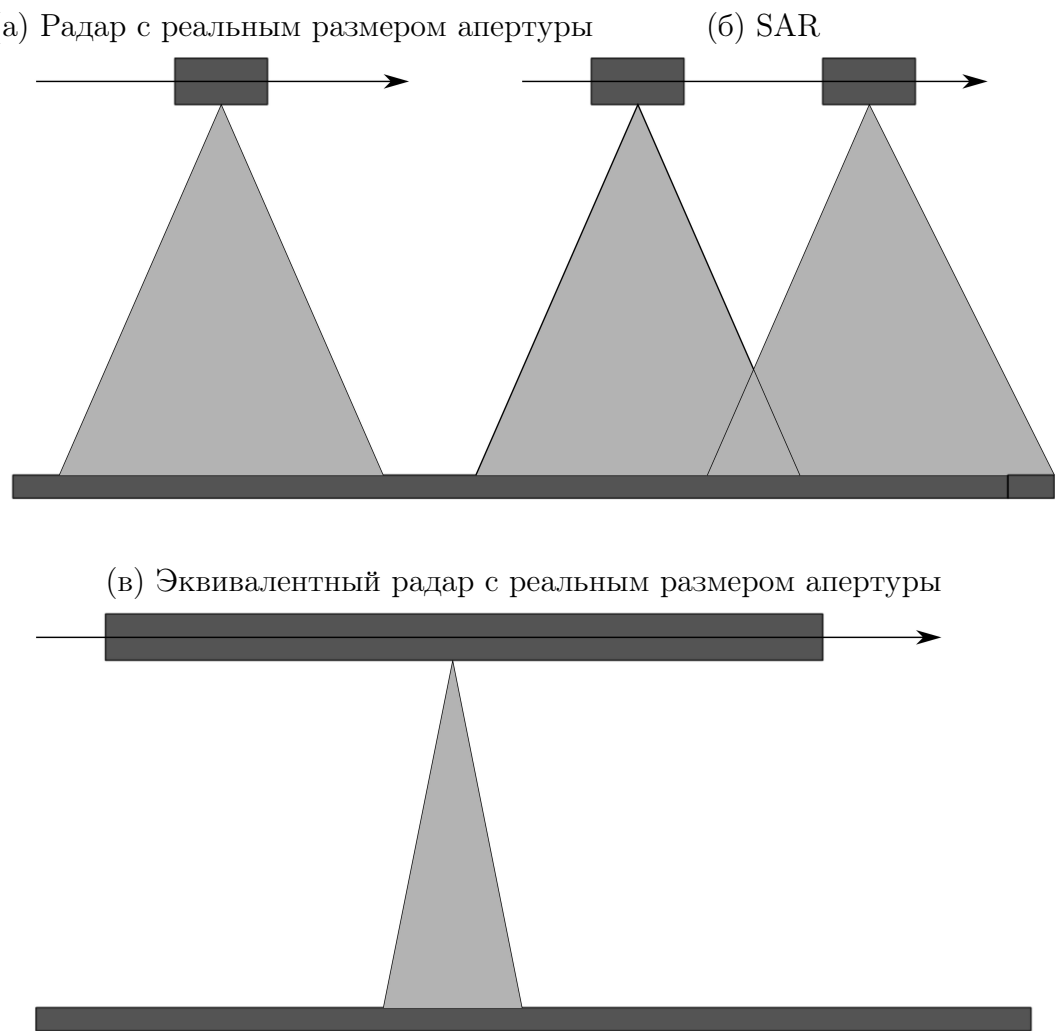


Рис. 2: Эквивалентность реальной и синтезированной апертуры

пространственное разрешение - 3x3 м. Получаемое изображение представляет квадратную сцену с длиной равной ширине полосы съемки. Уровень обработки данных в используемых изображениях - SCS (Single Look Complex Slant - одиночная наклонная съемка). Такие изображения формируются на основе одиночного сигнала, шаг пикселя равно отстоит по азимуту и углу падения, сохраняется информация о фазе, данные представлены в формате комплексных чисел, отсутствует геопривязка, изображение ориентировано перпендикулярно направлению движения радиолокационного аппарата.

SAR-интерферометрия

Для SAR-интерферометрии используется два (или более) SAR-снимков одного и того же участка земной поверхности. В результате комплексного поэлементного перемножения снимков формируется интерферограмма. Разность интерферограмм, полученных из различных пар SAR-снимков, позволяет определять малые смещения земной поверхности.

Принцип дифференциальной интерферометрии проиллюстрирован на рис. 3. Для реализации этого принципа необходимо иметь как минимум два SAR-снимка, полученных в разные моменты времени. Минимальный интервал между съемками определяется периодом повторения орбиты при движении спутников вокруг Земли и составляет несколько

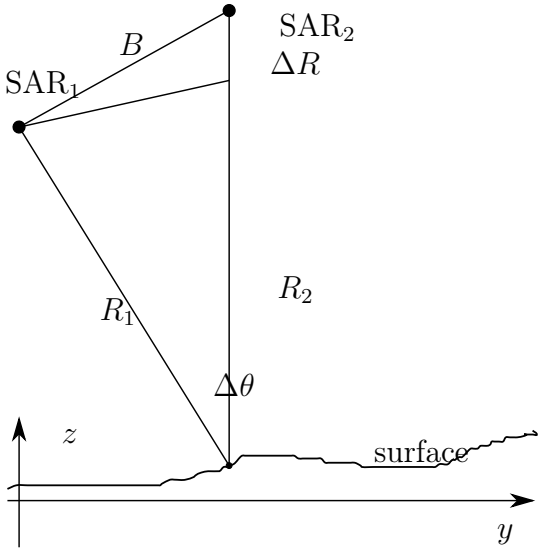


Рис. 3: Принцип SAR-интерферометрии

суток для многоспутниковой системы COSMO-SkyMed. При этом, не смотря на относительную стабильность орбиты, съемка одного и того же участка земной поверхности неизбежно осуществляется из разных положений, характеризующихся расстояниями R_1 и R_2. Важно, чтобы расстояние между двумя положениями спутника B, называемое *базовой линией* [1], не превышало нескольких сот метров. При большем значении базовой линии полностью пропадает корреляция между совмещаемыми фазовыми изображениями.

Опишем вкратце принцип интерферометрии. Отраженные волны u_1 и u_2 от точки поверхности x, принятые спутником в положениях R_1 и R_2, записываются в виде

$$u_1(R_1,x) = |u_1(R_1,x)|\exp\left(i\varphi_1(R_1,x)\right), \quad u_2(R_2,x) = |u_2(R_2,x)|\exp\big(i\varphi_2(R_2,x)\big).$$

Результат интерференции записывается в виде произведения:

$$v(\cdot) = u_1(\cdot)u_2^*(\cdot) = |u_1(\cdot)||u_2^*(\cdot)||\exp\big(i\Delta\varphi(\cdot)\big)|,$$

где

$$\Delta\varphi(\cdot) = \varphi_1(\cdot) - \varphi_2(\cdot)$$

— разность фаз.

Не принимая во внимание шумовые эффекты, запишем разность фаз $\Delta\varphi$ как функцию от разности хода ΔR:

$$\Delta\varphi = \frac{4\pi}{\lambda}\Delta R.$$

Таким образом, интерферограмма состоит из пикселов, каждому из которых соответствует разность фаз, зарегистрированных спутником, находящемся в различных положениях (на разных расстояниях от поверхности). Если эта разность фаз кратна 2π, то на интерферограмме наблюдается чередование темных и светлых полос. Каждая полоса характеризуется определенным значением разности фаз зарегистрированного сигнала при съемке с разных расстояний. В силу того, что съемка осуществляется из разных положений, существенный вклад в эту картину вносит рельеф территории (топографическая компонента). Скомпенсировав топографическую компоненту фазы можно получить интерферограмму, характеризующую смещения поверхности относительно первоначального

положения за время между съемками. Участки, не являющиеся устойчивыми отражателями радиолокационного сигнала, отображаются на интерферограмме в виде областей пикселов со случайными значениями разности фаз. Устойчивые отражатели радиолокационного сигнала демонстрируют относительно плавные изменения фазы на интерферограмме, а также зачастую характеризуются высокими значениями амплитудной компоненты. Изучив структуру полос на интерферограмме, можно выделить участки, на которых земная поверхность сдвинулась за время между съемками и оценить величину смещения.

Постановка задачи выделения устойчивых отражателей

Заданы изображения — матрицы размера $m \times n$ яркостей $\mathbf{Z}$, высот $\mathbf{H}$ и фаз $\mathbf{\Phi}$. Элементы z_{ij}, h_{ij}, φ_{ij} этих матриц принадлежат множеству $\mathbb{R}^1_+$.

Целочисленные величины — индексы i, j матриц поставлены в соответствие точкам x, y отрезков $\mathcal{X}, \mathcal{Y}$:

$$x = x(i), \quad y = y(j), \quad i \in \{1, ..., m\}, \quad j \in \{1, ..., n\}.$$

Соответствие задано таким образом, что величины x, y измеряются в метрах и интерпретируются в рамках данной задачи как координаты (x, y) некоторой картографической проекции, соответствующей изображениям $\mathbf{Z}, \mathbf{H}, \mathbf{\Phi}$.

Изображению, задаваемому матрицами $\mathbf{Z}, \mathbf{H}, \mathbf{\Phi}$ поставлены в соответствие функции яркости, высоты и фазы $Z, H, \Phi : \mathbb{R}^2 \to \mathbb{R}^1_+$.

Предполагается, что изображение $\mathbf{Z}$ содержит, в том числе, пятна (односвязные области, в которые можно вписать круг радиусом ξ), ассоциируемые с устойчивыми отражателями.

Требуется выделить на изображении множество отражателей, заданных координатами $(x, y) \in \mathbb{R}^2$, соответствующих однородным пятнам на изображении с резким изменением градиента яркости на границе. Требуется поставить в соответствие каждому отражателю (x, y) значение на фазовой составляющей $\varphi(x, y)$.

Алгоритм выделения устойчивых отражателей

Решается задача поиска устойчивых отражателей на амплитудной составляющей изображения $\mathbf{Z}$. На изображении отражатели представляются в виде однородных светлых пятен (в дальнейшем — блобов), обладающих свойством резкого изменения градиента на границе. Для решения задачи применяется LoG-детектор. Принцип, по которому строится LoG-детектор, состоит из двух последовательных этапов.

Свертка изображения с лапласианом гауссианы. Первый этап состоит в свертке яркостной составляющей изображения $Z(x, y)$ с лапласианом Δ гауссианы $G(x, y, \sigma)$, где

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\big(-\frac{x^2 + y^2}{2\sigma^2}\big),$$

$$\Delta G(x, y, \sigma) = \big(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\big) G(x, y, \sigma).$$

Введем также обозначение для нормированного на σ^2 лапласиана:

$$\hat{\Delta} G(x, y, \sigma) = \sigma^2 \Delta G(x, y, \sigma).$$

Свертка изображения записывается в виде

$$\hat{\Delta} Z(x, y, \sigma) = \iint \hat{\Delta} G(x - x', y - y', \sigma) Z(x, y) dx' dy'.$$

$\hat{\Delta}Z(x, y, \sigma)$ — результат свертки изображения $Z(x, y)$ с лапласианом гауссиана $\hat{\Delta}G(x, y, \sigma)$. Свертка выполняется для k различных значений параметра σ_t, $t \in \{1, ..., k\}$, и в результате образуется k отфильтрованных изображений $\hat{\Delta}Z(x, y, \sigma_t)$, $t \in \{1, ..., k\}$.

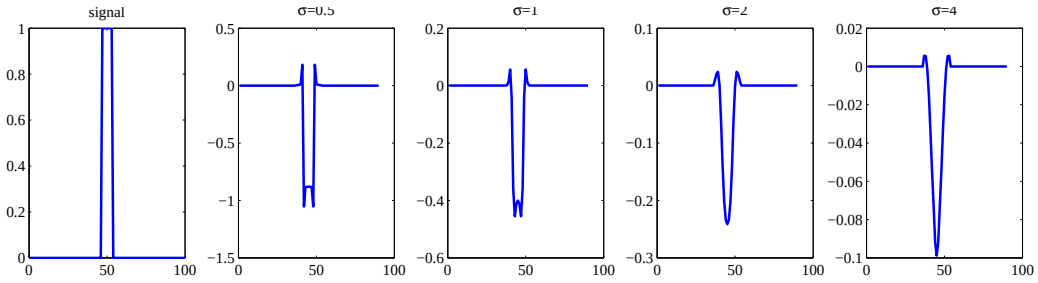


Рис. 4: Принцип LoG-детектора

Иллюстрация LoG-детектора показана на рис. 4 (рассматривается случай одномерного сигнала $Z(x)$). На крайнем левом графике показан входной сигнал $Z(x)$, который в наших терминах является блобом фиксированного размера. На последующих четырех графиках показана свертка $\hat{\Delta}Z(x, \sigma)$ исходного сигнала $Z(x)$ с лапласианом гауссианы $\hat{\Delta}G(x, \sigma)$ для различных значений σ. При малых значениях σ свертка $\hat{\Delta}Z(x, \sigma)$ имеет два локальных минимума и детектирует отдельно границы блоба. При увеличении σ происходит сглаживание сигнала и соединение двух локальных минимумов в один. При $\sigma = 2$ функция свертки имеет один ярко выраженный минимум, таким образом, значение параметра $\sigma = 2$ отвечает за характерный размер блоба.

Поиск характерного размера блоба. Для того, чтобы определить характерный размер блоба $\sigma_{\hat{t}}$ в точке (x, y), необходимо вычислить минимум функции

$$\hat{t} = \arg\min_t \hat{\Delta}Z(x, y, \sigma_t), \quad t \in \{1, ..., k\}.$$

Отметим, что если эта функция не имеет ярко выраженного минимума, или имеет несколько минимумов, то точка с координатами (x, y) не является центром блоба. Таким образом, на изображении Z метод находит некоторое разреженное множество координат центров блобов $\{A_i\}_{i=1}^N$ и их соответствующих размеров $\{\sigma_i\}_{i=1}^N$.

Аффинная адаптация размеров блоба. В этом параграфе под координатами (x, y) будем понимать координаты центров найденных блобов $\{A_i\}_{i=1}^N$.

Отметим, что найденными нами блобы характеризуются двумя параметрами: расположением (x, y) и радиусом σ. Пятна на изображении при этом могут быть произвольной формы, и для уточнения формы блоба предлагается произвести корректировку, или аффинную адаптацию, размеров блоба. Метод аффинной адаптации основывается на анализе вариации яркости изображения в зависимости от сдвига пиксела $\begin{bmatrix} x \\ y \end{bmatrix}$ в направлении вектора $\begin{bmatrix} u \\ v \end{bmatrix}$ (см. [10]):

$$E(u, v) = \sum_{x,y} w(x, y)\big(Z(x + u, y + v) - Z(x, y)\big)^2,$$

где $w(x, y)$ — функция окна (прямоугольного или гауссова). Для небольших сдвигов справедливо приближение:

$$E(u, v) \approx \begin{bmatrix} u & v \end{bmatrix} \mathbf{M} \begin{bmatrix} u \\ v \end{bmatrix},$$

где $\mathbf{M}$ — матрица, состоящая из взвешенных значений функции интенсивности:

$$\mathbf{M} = \sum_{x,y} w(x, y) \begin{bmatrix} Z_x^2 & Z_x Z_y \\ Z_y Z_x & Z_y^2 \end{bmatrix}.$$

Матрица $\mathbf{M}$ является диагональной и допускает разложение

$$\mathbf{M} = \mathbf{R}^{-1} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{R}.$$

Матрица $\mathbf{M}$ определяет уравнение эллипса, показанного на рис. 5, с направлением, задаваемым поворотом $\mathbf{R}$, и главными осями $\lambda_1^{-\frac{1}{2}}$, $\lambda_2^{-\frac{1}{2}}$. Форма этого эллипса описывает форму соответствующего блоба.

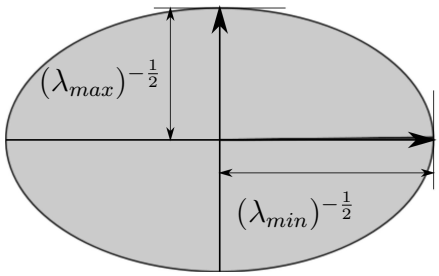


Рис. 5: Эллипс, соответствующий блобу.

Вычислительный эксперимент

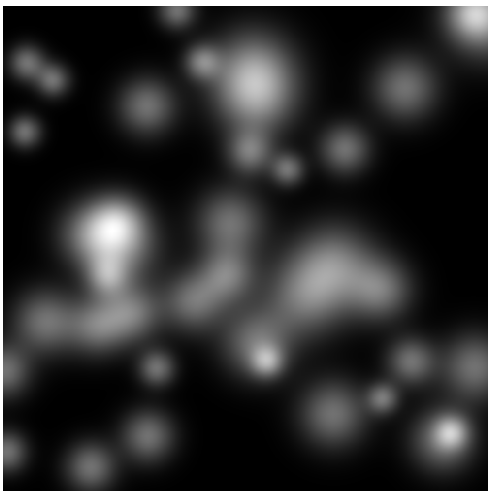
Эксперимент проводился на реальных и модельных данных.

В качестве модельных данных сгенерирована амплитудная составляющая изображения, состоящая из набора гауссиан разного масштаба, показанная на рис. 6 (a), и фазовая составляющая на рис. 6 (b). Предложенный выше алгоритм выделил на этой составляющей 24 блоба, что показано на рис. 6 (c).

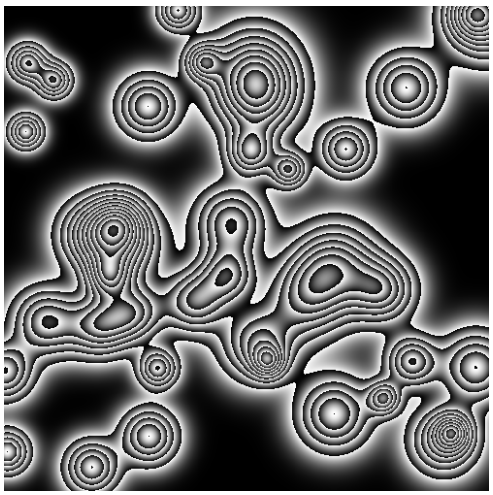
В качестве реальных данных использованы фрагменты снимков, полученных системой COSMO-SkyMed, размером 2000×2000. На рис. 7 (a) показана амплитудная составляющая снимка, на рис. 7 (b) — фазовая составляющая. На амплитудной составляющей алгоритм выделил 1970 блобов, соответствующих утойчивым отражателям, что показано на рис. 7 (c).

Заключение

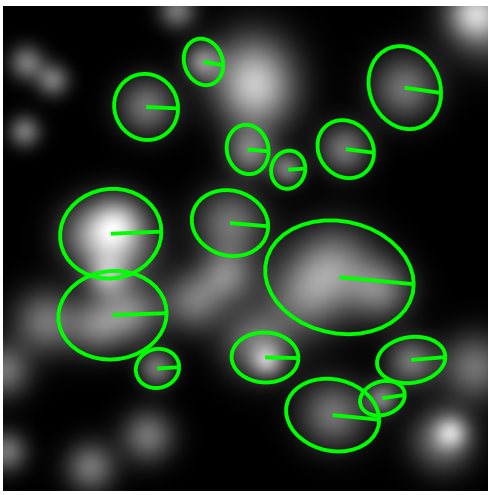
Исследована задача выделения системы устойчивых отражателей на SAR-снимках земной поверхности. Предложен метод нахождения устойчивых отражателей, основанный на широко распространенном в обработке изображений способе поиска блобов. Этот способ представляет собой свертку изображения с лапласианом гауссиан разных масштабов и поиск максимумов в полученной серии сглаженных изображений. Предложен алгоритм



(a) Амплитудное изображение

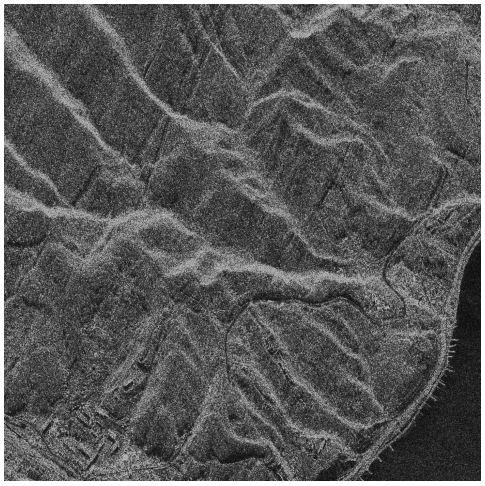


(b) Фазовое изображение

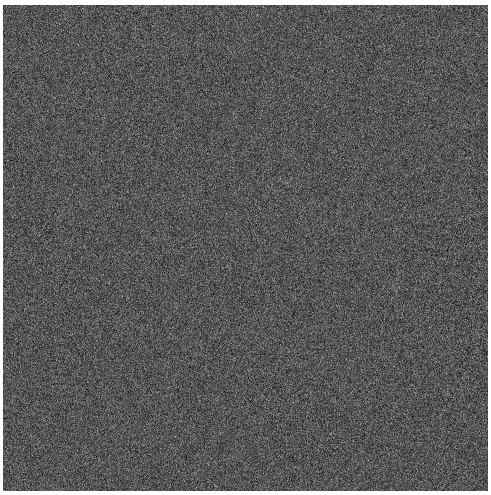


(c) Выделенные отражатели

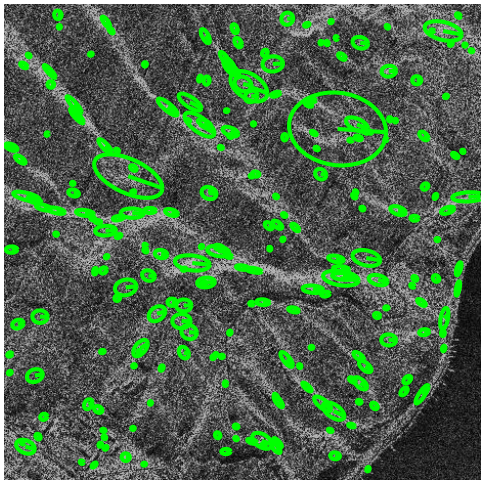
Рис. 6: Модельные данные



(a) Амплитудное изображение



(b) Фазовое изображение



(c) Выделенные отражатели

Рис. 7: Реальные данные.

уточнения эллипсоидальной формы блобов, основанный на сингулярном разложении матрицы, состоящей из взвешенных значений производной функции интенсивности. Адекватность работы алгоритма проиллюстрирована на синтетических и реальных данных.

Литература

[1] Hartl P., Bamler R. Synthetic aperture radar interferometry. *Inverse Problems*, 14:R1–R54, 1998.

[2] Harger R.O. Synthetic aperture radar fundamental and image processing. *EARSeL Advances in Remote Sensing*, 2:268–286, 1993.

[3] Costantini M. A new method for identification and analysis of persistent scatterers in series of SAR images. In *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, 2008.

[4] Rocca F., Ferretti A., Prati C. Permanent scatterers in SAR interferometry. *IEEE Transactions On Geoscience And Remote Sensing*, 39:8–20, 2001.

[5] Rocca F., Ferretti A., Prati C. Nonlinear subsidence rate estimation using permanent scatterers in differential sar interferometry. *Ieee Transactions On Geoscience And Remote Sensing*, 38:2202–2212, 2000.

[6] Agram P.S. *Persistent Scatterer Interferometry In Natural Terrain.* PhD thesis, Stanford University, 2012.

[7] Schmid C., Mikolajczyk K. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60:63–86, 2004.

[8] Lindeberg T. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[9] Italian Space Agency. *COSMO-SkyMed SAR Products Handbook.* http://www.e-geos.it/products/pdf/csk-product

[10] Бондаренко А.В., Ососков М.В., Моржин А.В., Визильтер Ю.В., Желтов С.Ю. *Обработка и Анализ Изображений в Задачах Машинного Зрения.* М.: Физматкнига, 2010.