На правах рукописи

ИОФИНА ГАЛИНА ВЛАДИМИРОВНА

ВЫБОР ОПТИМАЛЬНЫХ МЕТРИК В ЗАДАЧАХ РАСПОЗНАВАНИЯ С ПОРЯДКОВЫМИ ПРИЗНАКАМИ

05.13.17 — теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2010

Работа выполнена в Московском физико-техническом институте (государственном университете)

Научный руководитель:	доктор физико-математических наук, академик РАН **Юрий Иванович Журавлёв**
Официальные оппоненты:	доктор физико-математических наук **Олег Валентинович Сенько**
	кандидат физико-математических наук **Андрей Сергеевич Инякин**
Ведущая организация:	Научно-исследовательский институт системных исследований РАН

Защита диссертации состоится « ____ » _____ 2010 г. в _____ на заседании диссертационного совета Д 002.017.02 в Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан « ____ » _____ 2010 г.

Учёный секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор В. В. Рязанов

Общая характеристика работы

Актуальность темы. Мера близости в задачах анализа данных часто играет решающую роль. Поэтому ее выбору уделяется особое внимание. При решении задач распознавания часто производится предобработка данных, т. е. ставится задача выбора или построения оптимальных в том или ином смысле метрик или функций расстояния на объектах. Оптимальные функции близости выбираются из определенного семейства путем изменения параметров. Обычно, если в задачах классификации объекты заданы векторами признаков, то вначале на признаках фиксируется какая-нибудь стандартная функция расстояния (например, евклидова метрика). Далее на основе полученной информации формируется функция близости для объектов.

Однако имеет смысл поставить задачу оптимизации не столько функции близости на объектах, сколько на значениях признаков этих объектов. Тогда могут быть выделены преимущества или недостатки не только объекта в целом, но и каждого признака по-отдельности. Таким образом можно усиливать наиболее важные признаки и ослаблять шумовые, которые не должны влиять на распознавание, однако при использовании стандартной метрики играют существенную роль.

Кроме того, в литературе чаще всего рассматриваются задачи поиска оптимальных метрик (с выполненным неравенством треугольника) на действительных признаках. Из-за информатизации общества число задач с порядковыми признаками с каждым годом увеличивается. Поэтому изучение особенностей задач с порядковыми признаками и оптимизация метрик и функций расстояния на порядковых признаках как их важной части, становятся все более актуальными.

Цель работы — поиск и использование оптимальных функций расстояний, удовлетворяющих всем аксиомам метрик или полуметрик, кроме неравенства треугольника (которое заменено на условие порядка (расстояние между дальними значениями не меньше, чем между ближними) в первом случае и неравенство треугольника с операцией сумма по модулю натурального числа N во втором случае) в различных задачах распознавания образов с порядковыми признаками.

Научная новизна. Все результаты, полученные в диссертации, являются новыми. В работе впервые была исследована задача поиска функций расстояния в задачах распознавания с порядковыми признаками, среди функций от двух переменных, удовлетворяющих всем условиям метрики, кроме неравенства треугольника, замененного на условие порядка.

После определения структуры оптимальных функций расстояния была исследована их взаимосвязь с евклидовой метрикой. Это позволяет использовать найденные оптимальные функции расстояния при решении задач распознавания методами, работающими только для евклидовой метрики, а также в задачах с признаками различных типов.

Впервые для решения задач распознавания с порядковыми признаками предложено использовать функции расстояния, которые удовлетворяют всем аксиомам полуметрики, однако на области значений функции вместо обычного сложения используется сумма по модулю чисел от 1 до 7.

При изучении алгебраических структур алгоритмов вычисления оценок впервые рассматривалась задача, когда алгоритмы различались не значениями параметров, а функциями расстояния на признаках. Причем брались оптимальные найденные ранее функции расстояния. Впервые была поставлена зада-

ча поиска метрики, которая обеспечила бы *регулярность* задачи распознавания с порядковыми признаками (т. е. в задаче распознавания должны были выполняться три естественных условия: 1) множества эталонов каждого из классов попарно различны, 2) в контрольной выборке нет ни одной пары объектов, неразличимых относительно эталонов, 3) обучающая и контрольная выборки не пересекаются).

Однако из-за того, что области значений функций расстояния ограничены (тремя значениями), при фиксированных метриках часть информации об объектах теряется. В этом случае потерянную информацию было предложено заменить возможностью выбора метрик. Поэтому были сформулированы и решены задачи поиска критериев корректности линейного и алгебраического замыканий алгоритмов вычисления оценок при фиксировании всех стандартных параметров алгоритма, но при возможности выбирать функции расстояния из некоторого специально заданного множества или множества всевозможных рассматриваемых метрик.

Методы исследования. В исследовании использовались комбинаторные рассуждения, методы теории графов, линейной алгебры, оптимизации, теории сложности. При изучении корректности моделей АВО использовались методы и подходы алгебраического подхода к решению задач распознавания, разработанные Ю. И. Журавлёвым, К. В. Рудаковым, В. Л. Матросовым, А. Г. Дьяконовым и др. Эксперименты проводились с использованием программного продукта Matlab.

На защиту выносятся следующие результаты, полученные для задачи распознавания с порядковыми признаками:

1. Представлен метод поиска наилучшей функции расстояния,

удовлетворяющей условию порядка и не удовлетворяющей неравенству треугольника.

2. Предложены методы работы по использованию полученных оптимальных функций расстояния в задачах распознавания со смешанными признаками.
3. Найдены всевозможные полуметрики в пространстве с суммой по модулю N для $N = 1, 2, \ldots, 7$, и сформулирован ряд теорем, справедливых для произвольных N.
4. Исследовано влияние выбора метрик и функций расстояния на порядковых признаках в алгоритмах вычисления оценок на корректность алгебраического замыкания модели АВО.
5. Получены условия корректности линейного замыкания модели АВО и оценки минимальной степени, необходимой для корректности алгебраического замыкания модели АВО при возможности выбора метрик на признаках из множества оптимальных функций расстояния и метрик.

Теоретическая и практическая значимость. Работа носит, в основном, теоретический характер. Совокупность результатов, полученных в диссертации, показывает, что иногда полезно использовать оптимальные функции расстояния на признаках при невыполнении в обычном смысле всех аксиом метрик (в частности, немонотонные функции расстояния). Методы, представленные в диссертации, могут быть непосредственно использованы на практике, а также служить основой для дальнейших теоретических исследований метрик на признаках. Эффективность полученных алгоритмов подтверждена решением практических задач.

Апробация работы. Результаты работы неоднократно докладывались на научных семинарах К. В. Рудакова и на конферен-

циях:

— научные конференциии МФТИ, 2006–2009 гг. [1,3,8,12];
— всероссийских конференциях «Математические методы распознавания образов» ММРО-13 (2007) и ММРО-14 (2009) [2,11];
— международная конференция «Интеллектуализация обработки информации» ИОИ-7 (2008) [6].
— международные конференции студентов, аспирантов и молодых ученых «Ломоносов-2008», «Ломоносов-2010» [4,14];

Публикации. Результаты работы изложены в статье «Журнала вычислительной математики и математической физики» [13], двух статьях журнала «Pattern Recognition and Image Analysis» [10,15], двух статьях сборника «Моделирование процессов обработки информации» [5,9], статье журнала «Таврический вестник» [7], а также в трудах конференций [1–4,6,8,11,12,14] (всего 15 публикаций, из которых три из списка ВАК). Описания отдельных результатов работы включались в научные отчеты по проектам РФФИ, № 08-01-00636 и № 08-01-00405.

Структура и объём работы. Работа состоит из оглавления, введения, четырёх глав, заключения и списка литературы (82 пункта).

Общий объём работы — 105 стр.

Краткое содержание работы по главам

В автореферате сохранена нумерация основных утверждений (определений, лемм, теорем и их следствий), принятая в тексте работы. Нумерация формул сквозная.

Во введении дана общая характеристика работы, обоснована актуальность, определено направление исследований, даны обзоры исследований по теме диссертации и основных результатов.

Глава 1. Различные критерии оптимальности функций расстояния в алгоритмах классификации

В **§1.1** рассмотрена задача классификации с двумя классами. Дано множество объектов $\{S^1, \ldots, S^{m_1}\}$ из класса K_1 , и $\{S^{m_1+1}, \ldots, S^{m_1+m_2}\}$ из класса K_2. Каждый объект принадлежит признаковому пространству размерности n. Значения признаков принадлежат конечному множеству $E^N = \{0,1, \ldots, N-1\}$, в котором задано отношение порядка $0 \leqslant 1 \leqslant 2 \leqslant \ldots \leqslant N-1$.

На каждом признаке задана своя *функция расстояния*, которая удовлетворяет всем аксиомам метрики кроме неравенства треугольника. То есть функция $\rho(x, y) : E^N \times E^N \to E^M$ удовлетворяет следующим условиям:

1. $\rho(x, y) = 0 \Leftrightarrow x = y$,
2. $\rho(x, y) = \rho(y, x)$,
3. $x \geqslant y \Rightarrow \rho(x, z) \geqslant \rho(y, z),\ \rho(x, z) \geqslant \rho(x, y),\ \forall z \in E^N, z \leqslant y$.

Последнее условие будем называть *условием порядка*.

Так как множество определения функции $E^N \times E^N$ ограничено, то функцию $\rho(x, y)$ можно представить как матрицу попарных расстояний элементов множества E^M. Занумеруем зна-

чения функции расстояния на парах (r,t) следующим образом:

$$\begin{pmatrix} 0 & x_1 & x_2 & x_3 & \dots & x_{N-1} \\ x_1 & 0 & x_N & x_{N+1} & \dots & x_{2N-3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N-2} & x_{2N-4} & x_{3N-5} & \dots & \dots & x_{N(N-1)/2} \\ x_{N-1} & x_{2N-3} & x_{3N-6} & x_{4N-10} & \dots & 0 \end{pmatrix}$$

причём условие порядка записывается в виде:

$$\begin{aligned} &x_1 \leqslant \ldots \leqslant x_{N-1}, && x_2 \geqslant x_N, \\ &x_N \leqslant \ldots \leqslant x_{2N-3}, && x_3 \geqslant x_{N+1} \geqslant x_{2N-2}, \\ &\ldots, && \ldots, \\ &x_{N(N-1)/2-2} \leqslant x_{N(N-1)/2-1}, && x_{N-1} \geqslant \ldots \geqslant x_{\frac{N(N-1)}{2}}, \end{aligned}$$

$$x_k \in E^M \backslash \{0\}, k = 1, \ldots, \tfrac{N(N-1)}{2}.$$

Видно, что функция расстояния определяется $N(N-1)/2$ числами. Поэтому она представляется вектором размерности $N(N-1)/2$. Признаки считаются попарно независимыми, поэтому в этом параграфе ищется функция расстояния на одном признаке.

Если обозначить количество нулей, единиц, двоек, троек и так далее среди значений признака объектов из первого класса через $\xi_0, \xi_1, \ldots, \xi_{N-1}$ соответственно, то количество сравниваемых пар при попарном сравнении объектов из первого класса (r,t) можно представить в виде матрицы:

$$A_1 = \begin{pmatrix} \frac{\xi_0(\xi_0-1)}{2} & \xi_0\xi_1 & \xi_0\xi_2 & \dots & \xi_0\xi_{N-1} \\ \xi_0\xi_1 & \frac{\xi_1(\xi_1-1)}{2} & \xi_1\xi_2 & \dots & \xi_1\xi_{N-1} \\ \xi_0\xi_2 & \xi_1\xi_2 & \frac{\xi_2(\xi_2-1)}{2} & \dots & \xi_2\xi_{N-1} \\ \dots & \dots & \dots & \dots & \dots \\ \xi_0\xi_{N-1} & \xi_1\xi_{N-1} & \xi_2\xi_{N-1} & \dots & \frac{\xi_{N-1}(\xi_{N-1}-1)}{2} \end{pmatrix}$$

Здесь каждый элемент a_1^{rt} равен числу пар (r,t), встречающихся при попарном сравнении объектов из первого класса.

Аналогично, если количество нулей, единиц, двоек, троек и так далее среди значений признака объектов из второго класса обозначить через $\eta_0, \eta_1, \ldots, \eta_{N-1}$ соответственно, то матрицу A_2 можно записать как:

$$A_2 = \begin{pmatrix} \frac{\eta_0(\eta_0-1)}{2} & \eta_0\eta_1 & \eta_0\eta_2 & \ldots & \eta_0\eta_{N-1} \\ \eta_0\eta_1 & \frac{\eta_1(\eta_1-1)}{2} & \eta_1\eta_2 & \ldots & \eta_1\eta_{N-1} \\ \eta_0\eta_2 & \eta_1\eta_2 & \frac{\eta_2(\eta_2-1)}{2} & \ldots & \eta_2\eta_{N-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \eta_0\eta_{N-1} & \eta_1\eta_{N-1} & \eta_2\eta_{N-1} & \ldots & \frac{\eta_{N-1}(\eta_{N-1}-1)}{2} \end{pmatrix}$$

Здесь каждый элемент a_2^{rt} равен числу пар (r, t), встречающихся при попарном сравнении объектов из второго класса.

Матрицы A_1 и A_2 — это матрицы, характеризующие внутреннюю структуру классов K_1 и K_2 посредством данной выборки. Поэтому, данные матрицы можно назвать *матрицами коэффициентов внутриклассовых расстояний.*

Аналогично можно записать *матрицу коэффициентов межклассовых расстояний*:

$$B = \begin{pmatrix} \eta_0\xi_0 & \eta_0\xi_1 + \xi_0\eta_1 & \eta_0\xi_2 + \xi_0\eta_2 & \ldots & \eta_0\xi_{N-1} + \xi_0\eta_{N-1} \\ \eta_0\xi_1 + \xi_0\eta_1 & \eta_1\xi_1 & \eta_1\xi_2 + \xi_1\eta_2 & \ldots & \eta_1\xi_{N-1} + \xi_1\eta_{N-1} \\ \eta_0\xi_2 + \xi_0\eta_2 & \eta_1\xi_2 + \xi_1\eta_2 & \eta_2\xi_2 & \ldots & \eta_2\xi_{N-1} + \xi_2\eta_{N-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \eta_0\xi_{N-1} + \xi_0\eta_{N-1} & \eta_1\xi_{N-1} + \xi_1\eta_{N-1} & \eta_2\xi_{N-1} + \xi_2\eta_{N-1} & \ldots & \eta_{N-1}\xi_{N-1} \end{pmatrix}$$

Для формализации критериев оптимальности функции расстояния вводятся понятия *внутриклассовых* и *межклассового* расстояний как средних арифметических всех попарных расстояний внутри классов и между элементами из разных классов соответственно. Если обозначить внутриклассовое расстояние для класса K_1 через α_1, а для класса K_2 через α_2, то во введенных обозначениях можно получить:

$$\alpha_1 = 1/N_1 \sum_{\substack{r,t=1,\ldots,N-1 \\ r \neq t}} x_{rt}\xi_r\xi_t,$$

$$\alpha_2 = 1/N_2 \sum_{\substack{r,t=1,\ldots,N-1 \\ r \neq t}} x_{rt}\eta_r\eta_t,$$

где N_1 и N_2 — нормировочные множители.

Для межклассового расстояния, аналогично имеем

$$\beta = 1/N_0 \sum_{\substack{r,t=1,\ldots,N-1 \\ r \neq t}} x_{rt}(\xi_r \eta_t + \eta_r \xi_t),$$

где N_0 — нормировочный множитель.

Рассматривается критерий максимизации взвешенной разницы межклассового и среднего внутриклассового расстояний, т. е. максимизируется величина $\beta - 0.5\lambda(\alpha_1 + \alpha_2)$.

Таким образом, исходная задача может быть представлена в виде

$$\begin{cases} \beta - 0.5\lambda(\alpha_1 + \alpha_2) \to \max\limits_{x_k, k=1,\ldots,N(N-1)/2} \\ 1 \leqslant x_k \leqslant M-1, k = 1, \ldots, N(N-1)/2 \\ x_k \text{ — удовлетворяют условию порядка} \\ x_k \text{ — целое} \end{cases}$$

Здесь λ можно рассматривать как отношение весов межклассового и среднего внутриклассового расстояний соответственно, а $x_k = x_{rt}$, где

$$k = (N-1)+(N-2)+\cdots+(N-r)+t-r = (2N-1-r)r/2+t-r, \forall r \leqslant t.$$

Так как α_1, α_2, и β линейны по x_{rt}, по которым происходит оптимизация, то имеем задачу целочисленного линейного программирования:

$$\begin{cases} \sum\limits_{k=1}^{N(N-1)/2} \gamma_k x_k \to \max\limits_{x_k, k=1,\ldots,N(N-1)/2} \\ 1 \leqslant x_k \leqslant M-1, k = 1, \ldots, N(N-1)/2 \\ x_k \text{ — удовлетворяют условию порядка} \\ x_k \text{— целое} \end{cases}$$

где $\gamma_k = 1/N_0(\xi_r\eta_t + \eta_r\xi_t) - 0.5\lambda/N_1\xi_r\xi_t - 0.5\lambda/N_2\eta_r\eta_t$, $k = (2N - 1 - r)r/2 + t - r$, $\forall r \leqslant t$ — соответствующие коэффициенты.

Для данного критерия доказано, что для нахождения оптимальных функций расстояния достаточно рассматривать только матрицы, состоящие из чисел 1 и $M - 1$ и решать задачу линейного программирования (вместо задачи целочисленного программирования), т. е. справедлива

Теорема 1.1. Решением оптимизационной задачи

$$\begin{cases} \sum\limits_{k=1}^{N(N-1)/2} \gamma_k x_k \to \max\limits_{x_k, k=1,\ldots,N(N-1)/2} \\ min \leqslant x_k \leqslant max, k = 1, \ldots, N(N-1)/2, \\ x_k \text{ — удовлетворяют условию порядка} \end{cases}$$

для действительных x_k могут являться только векторы $b = (b_1, \ldots, b_{N(N-1)/2})$, в которых $b_k = min$ или $b_k = max$, $\forall k = 1, \ldots, \frac{N(N-1)}{2}$.

Для оценки сложности задачи получено число матриц расстояний размерности N:

Теорема 1.2. Число матриц функций расстояний размерности N, удовлетворяющих условию порядка, можно представить следующей рекуррентной формулой (через числа Каталана):

$$\begin{aligned} f(N) &= f(0)f(N-1) + f(1)f(N-2) + \ldots + f(N-1)f(0) = \\ &= \sum_{k=0}^{N-1} f(k)f(N-1-k), \\ f(0) &= 1, \quad f(1) = 1. \end{aligned}$$

Или в явной форме данную величину можно представить как: $f(N) = \frac{C_{2N}^N}{N+1}$.

В **§1.2** получено обобщение основных результатов предыдущего параграфа на случай l классов. Описание алгоритма поиска функций расстояния дано в матричных обозначениях. Считается, что все признаки имеют одинаковую размерность, и на

каждом признаке задана своя функция расстояния. Отдельно рассмотрен случай, когда для некоторого множества признаков функции расстояния совпадают.

В **§1.3** показан вариант применения найденных функций расстояния для решения задач распознавания с порядковыми признаками с помощью алгоритма вычисления оценок. Дополнительным критерием оптимальности взят функционал минимизации числа ошибок, совершаемых алгоритмом.

В АВО расстояния между объектами определяются с помощью функций близости. В работе используется следующая функция близости для объектов S_u и S_t.

$$B_{\varepsilon}^{\vec{\varepsilon}}(\tilde{\omega}S_u, \tilde{\omega}S_t) = \begin{cases} 1, & \text{если число невыполненных неравенств в системе} \\ & \{\rho_i(S_u^i, S_t^i) \leqslant \varepsilon_i, i = 1, \ldots, l\} \text{ не больше } \varepsilon; \\ 0, & \text{в противном случае,} \end{cases}$$

где $\tilde{\omega}$ — характеристический вектор, соответствующий опорному множеству (подмножеству множества признаков), ρ_i — метрика в множестве значений i-го признака, ε_i — точность измерения i-го признака, ε — минимальное число невыполненных неравенств.

В работе оптимизация происходила по функциям расстояния при фиксированных остальных параметрах алгоритма. Было введено понятие *эквивалентности метрических характеристик относительно задач распознавания*:

Определение 1.13. *Метрической характеристикой признака i* алгоритма АВО называется пара $\{\rho_i, \varepsilon_i\}$ (функция расстояния на признаке и соответствующий порог).

Определение 1.14. Будем говорить, что метрические характеристики признаков s и r *эквивалентны относительно задачи распознавания Z*, если при их использовании в АВО алгоритмы A_{ρ_s,ε_s} и A_{ρ_r,ε_r} дают одинаковые результаты для всех

объектов контрольной выборки, то есть для всех S_u, $u = 1, \ldots, q$ выполняются равенства $A_{\rho_s,\varepsilon_s}(S_u) = A_{\rho_r,\varepsilon_r}(S_u)$.

Если $\varepsilon_s = \varepsilon_r = \tilde{\varepsilon}$, то будем говорить об *эквивалентности метрик*.

Было показано, что справедлива следующая

Теорема 1.4. Пусть для решения задачи распознавания с порядковыми признаками Z при использовании АВО с фиксированными параметрами метрические характеристики $\{\rho_i, \varepsilon_i\}$, $i = 1, \ldots, n$ также фиксированы. Тогда справедливы следующие утверждения:

1. Если $\varepsilon_i = 0$ или $\varepsilon_i \geqslant M_i - 1$, то все метрики ρ_i эквивалентны относительно задачи распознавания Z.

2. Если $0 < \varepsilon_i < M_i - 1$, то существует метрика $\rho_i^*\colon E^{N_i} \times E^{N_i} \to \{0, 1, 2\}$ такая, что метрические характеристики $\{\rho_i, \varepsilon_i\}$ и $\{\rho_i^*, 1\}$ эквивалентны относительно задачи Z.

Таким образом, без ограничения общности, можно использовать только метрики, принимающие значения из множества $\{0, 1, 2\}$.

В **§1.4** рассмотрен случай, когда в задаче классификации признаки заданы на множестве $E^N = \{0, 1, \ldots, N-1\}$ с естественным отношением порядка $0 \leqslant 1 \leqslant \ldots \leqslant N-1$ и, кроме того, в пространстве E^N задана операция $\oplus$ — сложение по модулю N. Рассмотрена задача поиска всех полуметрик, то есть функций от двух аргументов, удовлетворяющих всем аксиомам полуметрик с операцией сумма по модулю N в неравенстве треугольника.

В начале параграфа дана постановка задачи, и получены утверждения, верные для случаев произвольных размерностей N. Далее по-отдельности рассмотрены случаи размерностей $t =$

1, 2, 3, 4, 5, 6, 7. Для каждого случая доказывалась теорема о том, что все матрицы, удовлетворяющие некоторым условиям, являются полуметриками; либо, что полуметрик, удовлетворяющих данным условиям, нет; либо давался метод определения является ли рассматриваемая матрица полуметрикой в данном пространстве.

Глава 2. Выбор функций расстояния в задачах распознавания со смешанными признаками

На практике часто встречаются задачи с признаками различных типов (или, используя терминологию Н.Г. Загоруйко, измеренными в различных шкалах). Известно, что номинальные шкалы являются наиболее слабыми, далее идут порядковые шкалы, наиболее мощные — количественные или интервальные шкалы. Ясно, что работа с математическими операциями в более слабых шкалах может оказаться некорректной или даже невозможной. Поэтому для использования стандартных алгоритмов анализа данных в задачах со смешанными признаками можно усиливать одни типы признаков или ослаблять другие.

В **главе 2** рассматриваются задачи распознавания со смешанными признаками. Главным образом, изучается случай преобразования порядковых признаков в действительные.

В **§2.1** получен общий вид матриц расстояний, соответствующих функциям расстояний с выполненным условием порядка, а также неравенством треугольника. Рассматриваемые матрицы имеют структуру, состоящую из единичных подматриц с нулевыми диагоналями и строк с элементами, равными $M-1$ (кроме нулевых диагональных элементов). Единичные подматрицы расположены вдоль диагонали. Такая структура матрицы по-

парных расстояний задает в евклидовом пространстве несколько правильных симплексов с единичными ребрами. Все вершины двух разных симплексов находятся на расстояниях $M-1$ друг от друга. Получено, что число таких матриц размерности N при $M \geqslant 4$ равно $f(N) = 6 \cdot 2^{N-3}$; при $M \leqslant 3$ равно $f(N) = \frac{C_{2 \cdot N}^{N}}{N+1}$.

В **§2.2** рассмотрена задача вложения найденных матриц расстояния $A = \{a_{ij}\}$ размером $N \times N$ в евклидово пространство размерности t. То есть решена задача поиска точек (векторов) в евклидовом пространстве размерности t, которые дают те же расстояния, что и значения порядковых признаков в матрице расстояний. На основе модели Торгенсона был сформулирован критерий возможности погружения.

В **§2.3** рассмотрена задача оценки размерности евклидова пространства, в которое можно поместить объекты задачи распознавания, располагающиеся на заданных матрицей $A = \{a_{ij}\}$ расстояниях. Таким образом, объекты помещались в евклидово пространство на расстояниях равных единице либо $M-1$ друг от друга, и так, чтобы они образовывали кластерную структуру.

Для случаев малых размерностей $t = 1, 2, 3$ были построены все возможные конфигурации. Для случая евклидова пространства произвольной размерности t было найдено максимальное число объектов, которые можно в него поместить при выполнении всех условий на расстояния.

Теорема 2.4. Пусть в матрице попарных расстояний A недиагональные элементы принимают значения из множества $\{1, M-1\}$, и выполнены все аксиомы метрики и условие порядка. Тогда, если точки можно поместить в пространство размерности t согласно матрице попарных расстояний A, то число точек не может превышать $2t$.

Из данной теоремы можно получить ограничение на размерность пространства, в которое можно поместить N точек, находящихся на заданных рассматриваемых расстояниях:

Следствие 2.1. Размерность пространства t, требуемая для размещения N произвольных точек на расстояниях, удовлетворяющих условию порядка, находится в интервале $(\lceil N/2 \rceil, N-1]$.

Глава 3. Выбор метрик в алгебраических замыканиях модели АВО

В главе 3 рассматриваются различные варианты использования метрик и функций расстояний с условием порядка в алгебраических замыканиях модели алгоритмов вычисления оценок.

В **§3.1** дана постановка задачи и основные определения, введённые Ю.И. Журавлёвым, В.Л. Матросовым, А.Г. Дьяконовым и другими исследователями. Приведены некоторые результаты, полученные ранее другими авторами и используемые в дальнейшем.

В **§3.2** найдены условия регулярности задачи распознавания с порядковыми признаками при выборе фиксированных метрик на признаках. Рассуждения построены на понятии *полной разделимости* объектов по признаку, т. е. возможности каждому значению признака взаимнооднозначно сопоставить строку подматрицы $\vec{R} = \Big(\big(\rho_1(S^j, S_i), \ldots, \rho_n(S^j, S_i)\big)\Big)_{i=1}^{q}{}_{j=1}^{m}$, соответствующую этому признаку. Была получена

Теорема 3.3. Пусть в задаче осуществлено деление на подгруппы по полностью делящимся признакам. Тогда задача регулярна (а модель $U(B^*)$ корректна) тогда и только тогда, когда на множествах $s_1 \times t_1, \ldots, s_r \times t_r$ неполностью делящихся при-

знаков можно задать метрику так, чтобы в каждой подгруппе были различные векторы (здесь s_j, t_j — числа различных значений j-го признака в объектах из обучающей и контрольной выборок).

Рассмотрены как случай произвольных метрик на признаках, так и случай функций расстояния, удовлетворяющих условию порядка (являющихся также метриками из-за ограниченности множеством $\{0, 1, 2\}$ её значений). Условия регулярности согласно критерию корректности были эквивалентны условию корректности задач распознавания.

Из-за того, что множество значений функции расстояния ограничено, часть информации об объектах теряется. В этом случае потерянную информацию предлагается заменить возможностью выбирать метрики. В данном случае понятие регулярности задачи некорректно, так как оно в своём определении использует фиксированную метрику во втором условии. Поэтому в **§3.3** сформулирована и решена задача поиска критерия корректности алгебраического замыкания алгоритмов вычисления оценок при фиксации всех стандартных параметров алгоритма, но при возможности выбора метрик из некоторого множества.

Поэтому для работы с задачами при варьировании метрик предлагается рассматривать задачи, удовлетворяющие первому условию регулярности, а также *неполному второму условию при выборе метрик* — отсутствию в контрольной выборке одинаковых объектов.

Получен критерий корректности алгебраического замыкания АВО при условии, что все стандартные параметры алгоритма фиксированы, и пара алгоритмов в замыкании различается только используемыми в них метриками, которые выбираются из некоторого заданного множества метрик.

В **§ 3.4** рассмотрен частный случай алгебраического замыкания — алгебраическое замыкание наименьшей степени — линейное замыкание модели ABO. Получены критерии корректности линейного замыкания ABO при возможности выбора функций расстояния на порядковых признаках. Отдельно рассмотрены случаи задач распознавания с непересекающимися и пересекающимися классами, а также при выборе метрик из произвольного множества допустимых метрик (со значениями из ограниченного множества $\{0, 1, 2\}$) и при выборе метрик из множества функций расстояния, удовлетворяющих условию порядка.

В **§ 3.5** получена оценка минимальной степени корректного алгебраического замыкания ABO при возможности выбора функций расстояния на признаках. В случае метрик минимальная степень равняется 1, т. е. корректно уже линейное замыкание, а в случае функций расстояния, удовлетворяющих условию порядка, степень алгебраического замыкания не может превышать двух.

Глава 4. Эксперименты

Для определения целесообразности использования полученных функций расстояния на признаках был поставлен ряд экспериментов. Проведено сравнение использования в задачах распознавания различных оптимальных функций расстояний (с заменой неравенства треугольника в определении метрики условием порядка или в полуметрике неравенством треугольника с суммой по модулю некоторого числа) с использованием евклидовой метрики (манхэттенское расстояние) или расстояния Хэмминга.

В **§ 4.1** представлена серия модельных примеров, сравнива-

ющих качество работы алгоритма ближайшего соседа при использовании различных функций расстояния. В **§ 4.2** показаны результаты экспериментов с реальными задачами из репозитория UCI: breast cancer wisconsin, car, iris, wine, glass, ionosphere.

В **заключении** подводится краткий итог полученных результатов диссертации.

Публикации по теме диссертации

[1] *Иофина Г. В.* Выбор наилучшей метрики в алгоритме распознавания по ближайшему соседу // Сборник трудов 49-й научной конференции МФТИ. — М.: МФТИ, 2006. — С. 266–267.

[2] *Иофина Г. В., Кропотов Д. А.* Поиск оптимальной метрики в задачах классификации с порядковыми признаками // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 137–140.

[3] *Иофина Г. В.* Оптимизация метрик для порядковых признаков в задачах распознавания // Сборник трудов 50-й научной конференции МФТИ. — М.: МФТИ, 2007. — С. 98–100.

[4] *Иофина Г. В., Ветров Д. П., Кропотов Д. А.* Восстановление объектов в евклидовом пространстве по оптимальной метрике в задаче распознавания образов с порядковыми признаками // Ломоносов-2008: Материалы конф. — М.: Издательский отдел факультета ВмиК МГУ, 2008. — С. 38.

[5] *Иофина Г. В., Ветров Д. П., Кропотов Д. А.* Восстановление объектов в евклидовом пространстве по оптимальным матрицам близости // Моделирование процессов обработки информации. — М.: МФТИ, 2008. — С. 110–114.

[6] *Иофина Г. В.* Многомерное шкалирование в случае матриц попарных расстояний с элементами из конечного множества //

Интеллектуализация обработки информации: Тезисы докл. — Симферополь, 2008. — С. 112–113.

[7] *Иофина Г. В.* Многомерное шкалирование в случае матриц попарных расстояний с элементами из конечного множества // Таврический вестник информатики и математики. — 2008. — № 1. — С. 223–230.

[8] *Иофина Г. В.* Оптимизация алгоритмов вычисления оценок по метрикам на признаковых описаниях объектов // Сборник трудов 51-й научной конференции МФТИ. — М.: МФТИ, 2008. — С. 39–42.

[9] *Иофина Г. В.* Эквивалентные метрики в алгоритмах вычисления оценок в задачах с порядковыми признаками // Моделирование процессов обработки информации. — М.: МФТИ, 2009. — С. 65–69.

[10] *G. V. Iofina* Optimal Metrics in Classification Problems with Ordered Features and an Arbitrary Number of Classes // **Pattern Recognition and Image Analysis**. — 2009. — Vol. 19, no. 2. — Pp. 284–288.

[11] *Иофина Г. В.* Критерии корректности алгебраического замыкания модели АВО в задачах с порядковыми признаками // Докл. всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 37–41.

[12] *Иофина Г. В.* Критерии корректности алгебраического замыкания модели АВО при варьировании метрик на признаках // Сборник трудов 52-й научной конференции МФТИ. — М.: МФТИ, 2009. — С. 67–70.

[13] *Иофина Г. В.* Поиск оптимальной метрики в задачах классификации с порядковыми признаками // **ЖВМ и МФ**. — 2010. — Т. 50, № 3. — С. 585–592.

[14] *Иофина Г. В.* Условия корректности линейного замыкания модели АВО // Ломоносов-2010: Материалы конф. — М.: Издатель-

ский отдел факультета ВмиК МГУ, 2010. — С. 85–86.

[15] *G. V. Iofina* A Study of Metrics in Finite Sets for Application in Classification and Recognition Problems // **Pattern Recognition and Image Analysis**. — 2010. — Vol. 20, no. 4. — Pp. 238–246.