На правах рукописи

Зухба Анастасия Викторовна

ОЦЕНКА ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ ЗАДАЧ ОТБОРА ЭТАЛОННЫХ ОБЪЕКТОВ И ПРИЗНАКОВ

Специальность 01.01.09 —
«Дискретная математика и математическая кибернетика»

Автореферат

диссертации на соискание учёной степени
кандидата физико-математических наук

Долгопрудный — 2018

Работа выполнена на кафедре интеллектуальных систем факультета управления и прикладной математики Московского физико-технического института (государственного университета)

Научный руководитель:
Воронцов Константин Вячеславович, доктор физико-математических наук, профессор РАН

Официальные оппоненты:
Еремеев Антон Валентинович, доктор физико-математических наук, доцент, Омский филиал Федерального государственного бюджетного учреждения науки Института математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Лаборатория дискретной оптимизации, ведущий научный сотрудник

Михальский Анатолий Иванович, кандидат технических наук, доктор биологических наук, старший научный сотрудник, ФГБУН Институт проблем управления им. В.А. Трапезникова Российской академии наук, Лаборатория № 38 «Управления по неполным данным», главный научный сотрудник

Ведущая организация: Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук».

Защита состоится _____ 2018 года в ____ часов ____ минут на заседании диссертационного совета Д 212.156.05, созданного на базе Московского физико-технического института (государственного универсиета), по адресу: 141700, Московская обл., г. Долгопрудный, Институтский пер., д. 9, ауд. 903 КПМ.

С диссертацией можно ознакомиться в библиотеке Московского физико-технического института (государственного университета) и на сайте университета http://www.mipt.ru/.

Автореферат разослан «____»_____ 2018 года.

Ученый секретарь
диссертационного совета Федько Ольга Сергеевна

Общая характеристика работы

Диссертационная работа посвящена проблеме отбора объектов и признаков для построения метрических и монотонных классификаторов. В работе предложены оптимизационные постановки задач отбора объектов и признаков, произведена оценка их вычислительной сложности. Предложен приближенный алгоритм монотонизации обучающей выборки с одновременным отбором объектов и признаков. Произведена экспериментальная проверка алгоритма на данных задачи медицинской диагностики.

Актуальность темы исследования и степень её разработанности

Широкое распространение компьютерных технологий для сбора и накопления данных практически во всех областях жизни стимулирует развитие методов интеллектуального анализа данных, предсказательного моделирования, машинного обучения. В частности, для автоматизации принятия решений широко используются модели, методы и алгоритмы распознавания образов или классификации. На практике к этим методам предъявляются требования высокой обобщающей (предсказательной) способности и низкой вычислительной сложности.

Задача классификации заключается в том, чтобы по заданному конечному множеству объектов, разделенных на классы, построить функцию классификации, которая произвольному объекту той же природы ставит в соответствие один из заданных классов. Например, в задачах медицинской диагностики объектами являются признаковые описания состояния человека; в роли признаков могут выступать симптомы, результаты обследований или биомаркеры; классы соответствуют диагностируемым заболеваниям.

Большинство моделей классификации явно или неявно формализуют гипотезу компактности — неформальное предположение о том, что схожие объекты чаще принадлежат одному классу, чем разным. Успех в решении задачи классификации во многом зависит от того, насколько адекватно применяемая модель классификации выражает понятие «сходства» объектов.

При решении практических задач классификации часто возникает необходимость отбора объектов и признаков. Отбор признаков (feature selection, FS) необходим для выявления информативных подпространств признаков, в которых выполняется гипотеза компактности. Отбор объектов (prototype selection, PS) необходим для отсева ошибочных объектов (выбро-

сов) и выявления типичных представителей классов (эталонов), достаточных для понимания структуры класса и надёжной классификации остальных объектов. Кроме того, отбор как объектов, так и признаков, позволяет сокращать объём хранимых данных, уменьшать время обучения алгоритма, повышать обобщающую способность и устойчивость классификации. Задачи отбора объектов и признаков в литературе, как правило, рассматриваются по отдельности. Редкое исключение составляют работы новосибирской школы распознавания образов (Н. Г. Загоруйко, Г. С. Лбов, И. А. Борисова, В. В. Дюбанов, О. А. Кутненко и др.). Целесообразность единого алгоритмического решения этих двух задач следует из того факта, что результат отбора объектов может зависеть от признакового пространства, а результат отбора признаков может зависеть от способа отсева выбросов или выделения эталонов.

Возможность отбора эталонов наиболее естественно возникает в метрических методах классификации, поскольку для них понятие «сходства» объектов формализуется в явном виде. Примером удачной эвристики отбора эталонов для метрических классификаторов является метод FRiS-Stolp, предложенный Н. Г. Загоруйко. Альтернативный метод отбора эталонов — путём минимизации функционала полного скользящего контроля — был предложен М. Н. Ивановым и К. В. Воронцовым (2009). В этом методе явная оптимизация обобщающей способности позволяет улучшать множество отбираемых эталонов. Во всех перечисленных методах применяются жадные стратегии, не гарантирующие оптимальности решения, то есть что множество эталонов будет иметь минимальную мощность и/или обеспечивать минимальное значение критерия качества. Многие подходы не предполагают постановку задачи как оптимизационной. Вместо этого предлагается алгоритм отбора эталонов, качество которого исследуется «пост фактум» чисто эмпирически. Оптимизационные постановки задачи отбора эталонов и признаков, а также вопросы их вычислительной сложности, являются относительно малоисследованными.

В данной работе рассматриваются оптимизационные постановки задач отбора объектов и признаков в метрических классификаторах. Кроме того, данная задача естественным образом обобщается на случай монотонных классификаторов. Предположение о монотонности функции классификации возникает во многих прикладных задачах, и его явный учёт позволяет по-

вышать обобщающую способность метода классификации. В данной работе рассматривается полный набор оптимизационных постановок задач отбора объектов и признаков для построения монотонных классификаторов.

На практике часто пользуются линейными моделями классификации с неотрицательными коэффициентами, как самым простым способом реализации монотонного классификатора. Преимущество линейной модели — в её простоте и наличии готовых реализаций. Однако для многих задач линейная модель представляется слишком жёсткой. Множество монотонных функций существенно шире, чем множество линейных функций с неотрицательными коэффициентами. Поэтому нелинейные монотонные модели классификации имеют преимущества в задачах со сложной разделяющей поверхностью, что было показано различными авторами на примерах задач медицинской диагностики, ранжирования поисковой выдачи, категоризации текстов, при построении композиций классификаторов.

Большинство методов монотонной классификации требуют предварительной монотонизации обучающей выборки. Монотонизация сводится к отбрасыванию объектов обучающей выборки, нарушающих условие монотонности и поиску пространства признаков, в котором условия монотонности выполняются. Сложность построения монотонного классификатора оценивается снизу вычислительной сложностью задачи монотонизации выборки.

Как и в случае с метрическими алгоритмами, вопросы оптимизационной постановки задачи монотонизации и их вычислительной сложности являются малоисследованными.

Цели и задачи

Целью данной работы является выяснение статуса вычислительной сложности задач отбора эталонных объектов и признаков для метрических и монотонных классификаторов. Для достижения поставленной цели решаются следующие задачи:

1. Оценить вычислительную сложность оптимизационных постановок задач отбора объектов и признаков для алгоритма ближайшего соседа.
2. Оценить вычислительную сложность оптимизационных постановок задачи монотонизации обучающей выборки.

3. Разработать алгоритм монотонизации с одновременным отбором объектов и признаков.

Научная новизна

1. Получена оценка вычислительной сложности задачи отбора объектов и признаков для алгоритма ближайшего соседа.
2. Предложена систематизация оптимизационных постановок задачи монотонизации выборки.
3. Получена оценка вычислительной сложности задачи монотонизации выборки.
4. Предложен и протестирован экспериментально алгоритм монотонизации выборки с одновременным отбором объектов и признаков.

Теоретическая и практическая значимость

Оптимизационная постановка позволяет получить оценки вычислительной сложности задач отбора объектов и признаков при различных целевых функциях и ограничениях. Систематизация получаемых задач дискретной оптимизации позволяет выбирать целевые функции, которые соответствуют решаемой прикладной задаче. Доказательство NP-полноты обосновывает применение субоптимальных эвристических методов для решения соответствующих оптимизационных задач. Предложенный в работе приближенный алгоритм монотонизации с одновременным отбором объектов и признаков частично решает проблему «застревания» в локальных минимумах, связанных с шумовыми объектами и неинформативными признаками. Для единственной постановки задачи, имеющей полиномиальную сложность, указан точный эффективный алгоритм решения.

Методология и методы

Для анализа постановок задач отбора объектов и признаков использовались элементы комбинаторики и теории графов. При оценке вычислительной сложности использовались методы сведения классических задач дискретной оптимизации (задачи о биклике, задачи о покрытии множеств подмножествами, задачи о вершинном покрытии) к задачам отбора объектов и признаков при обучении классификации. В целях проверки предложенного алгоритма монотонизации был проведен вычислительный эксперимент на прикладной задаче информационного анализа электрокардиосигналов.

Положения, выносимые на защиту

1. Оценка вычислительной сложности задачи отбора признаков и эталонных объектов для алгоритма ближайшего соседа.
2. Оценки вычислительной сложности задачи монотонизации обучающей выборки в различных постановках.
3. Приближенный алгоритм монотонизации обучающей выборки с одновременным отбором объектов и признаков.

Степень достоверности и апробация результатов

Достоверность теоретических результатов обеспечивается математическими доказательствами теорем. Результаты экспериментов соответствуют результатам, полученным другими авторами.

Основные результаты работы докладывались на следующих научных конференциях:

– 52-я, 53-я научные конференции МФТИ, 2009 [1], 2010 [2] Москва–Долгопрудный.
– Всероссийские конференции «Математические методы распознавания образов», ММРО-15, Петрозаводск 2011 [3]; ММРО-16, Казань 2013 [4]; ММРО-17, Светлогорск 2015 [5; 6]; ММРО-18, Таганрог 2017 [7].

Публикации

Основные результаты по теме диссертации изложены в десяти публикациях [1–10], две из которых опубликованы в изданиях из перечня, рекомендованного ВАК [8; 9], семь — в тезисах докладов [1–7].

Научная работа соискателя по теме диссертации была поддержана грантом РФФИ в рамках проекта №14-07-31240 мол_а.

Личный вклад автора в публикации с соавторами заключался в разработке и обосновании различных версий алгоритмов монотонизации и подготовке текста публикации.

Структура и объем диссертации Диссертация состоит из введения, четырех глав, заключения, списка обозначений, списка литературы и приложения. Список литературы содержит 54 наименования. Общий объем диссертационной работы 113 страниц.

Основное содержание работы

Во **введении** к диссертации обсуждаются: актуальность темы исследования, степень разработанности темы, цели и задачи, научная новизна, теоретическая и практическая значимость работы, методология и методы исследования, положения, выносимые на защиту, степень достоверности результатов, апробация результатов, публикации автора по теме диссертации, личный вклад автора в публикациях с соавторами.

В первой главе вводятся необходимые понятия и обозначения, определяются критерии качества классификаторов, рассматривается оптимизационная постановка обучения классификатора по прецедентам.

Пусть имеется множество объектов $\mathbb{X}$ и множество ответов Y, и существует функция $y\colon \mathbb{X} \to Y$, значение для которой известно только на конечном подмножестве $\{x_1, \ldots, x_L\} \subset \mathbb{X}$. Известные пары «объект–ответ» (x_i, y_i) называют *прецедентами*, а совокупность $X^L = (x_i, y_i)_{i=1}^L$, где $y_i = y(x_i)$, — *обучающей выборкой*.

Задача обучения по прецедентам состоит в том, чтобы построить отображение $\gamma\colon \mathbb{X} \to Y$, которое аппроксимирует функцию $y(x)$. Отображения $\gamma\colon \mathbb{X} \to Y$, аппроксимирующее функцию $y(x)$, называют *алгоритмами*.

Моделью алгоритмов называется параметрическое семейство отображений $\Gamma_\Theta = \{\gamma_\Theta(x,\theta) | \theta \in \Theta\}$, где $\gamma_\Theta\colon \mathbb{X} \times \Theta \to Y$, а Θ — множество допустимых значений параметров θ.

Методом обучения называется отображение μ, которое произвольной конечной выборке X^L ставит в соответствие алгоритм γ из Γ_Θ. Как правило, метод обучения μ сводится к решению оптимизационной задачи выбора параметра θ из множества Θ.

Индикатором ошибки алгоритма γ на объекте x_i называется функция, принимающая значение 0, если ответ алгоритма совпадает с истинным ответом, и 1 в противном случае:

$$I(x_i, \gamma(x_i)) = \big[y(x_i) \neq \gamma(x_i)\big].$$

Частота ошибок алгоритма γ на выборке X^L определяется как

$$\nu(\gamma, X^L) = \frac{1}{L}\sum_{i=1}^{L} I(x_i, \gamma(x_i)).$$

Если частота ошибок $\nu(\gamma, X) = 0$, говорят что алгоритм γ корректно работает на множестве объектов X.

Частота ошибок является одной из самых простых целевых функций, используемых при обучении алгоритмов. Ее недостатком является то, что малая частота ошибок на обучающей выборке еще не гарантирует, что построенный алгоритм будет также редко ошибаться на новых (контрольных) объектах.

Алгоритм обучения обладает *обобщающей* способностью (generalization ability), если вероятность ошибки на контрольных объектах не сильно отличается от ошибки на обучающей выборке. Если вероятность ошибки обученного алгоритма на контрольных объектах оказывается существенно выше, чем средняя ошибка на обучающей выборке, то говорят, что произошло *переобучение*. Для оценивания обобщающей способности метода используют функционалы качества, основанные на принципе скользящего контроля.

Пусть дана выборка X^L длинны L. Разобьём её на два непересекающихся подмножества: обучающую подвыборку (training set) X^ℓ и контрольную подвыборку (testing set) X^k, где $L = k + \ell$. Обозначим через (X_n^ℓ, X_n^k), $n = 1, \ldots, N$ всевозможные разбиения выборки X^L на обучающую и контрольную подвыборки, $N = C_L^\ell$.

Функционалом полного скользящего контроля (complete cross validation, CCV) называется средняя частота ошибок на контрольных подвыборках:

$$Q_k(\mu) = \frac{1}{N}\sum_{n=1}^{N} \nu(\mu(X_n^\ell), X_n^k).$$

Признаком f называют отображение $f : \mathbb{X} \to D_f$, где D_f — множество допустимых значений признака. Пусть дан набор признаков $\mathbb{F} = \{f_1, \ldots f_n\}$. Вектор $\big(f_1(x), \ldots f_n(x)\big)$ называют *признаковым описанием* объекта x. Как правило, объекты задаются своими признаковыми описаниями.

Отбор признаков (FS) и отбор объектов (PS) по критериям минимума частоты ошибок или полного скользящего контроля являются задачами дискретной оптимизации. В последнем разделе первой главы перечисляются известные факты о задачах дискретной оптимизации, которые используются в данной работе.

Вторая глава посвящена задачам отбор эталонов и признаков без ограничений монотонности для метрических алгоритмов. Метрическими называются алгоритмы классификации, основанные на измерении сходства между объектами. Типичным представителем метрических алгоритмов является алгоритм ближайшего соседа.

Метод обучения μ для классификации по ближайшему соседу (Nearest Neighbor classifier, NN) сводится к тривиальному запоминанию обучающей выборки. После этого произвольный классифицируемый объект $u \in \mathbb{X}$ относится к тому классу, которому принадлежит ближайший к нему обучающий объект. Для формализации понятия близости (сходства) на $\mathbb{X}$ вводится функция расстояния $\rho(x,x')$, вообще говоря, не обязательно метрика.

Для произвольного $x_i \in X^L$ положим $x_i \equiv x_{i0}$ и обозначим через $x_{i0}, x_{i1}, \ldots, x_{i,L-1}$ последовательность всех объектов выборки X^L, упорядоченную по возрастанию расстояний $\rho(x_i, x_{ij}),\ j = 0, \ldots, L-1$.

Обозначим через $r_m(x_i)$ ошибку, возникающую при замене известной классификации объекта x_i на ответ $y(x_{im})$ на m-ом соседе, то есть $r_m(x_i) = I(x_i, y(x_{im}))$.

Профилем компактности выборки X^L называется функция $P(m)$, выражающая долю объектов выборки, для которых правильный ответ не совпадает с правильным ответом на m-ом соседе:

$$P(m) = \frac{1}{L} \sum_{i=1}^{L} r_m(x_i); \quad m = 1, \ldots, L-1.$$

Профиль компактности $P(m)$ является формальным количественным выражением гипотезы компактности. Низкие, близкие к нулю, значения профиля компактности $P(m)$ для первых нескольких значений m означают, что гипотеза компактности на данной выборке выполняется.

Следующая теорема позволяет эффективно вычислять функционал полного скользящего контроля для метода ближайшего соседа.

Теорема 2.1.1. (Воронцов, 2004) Для задачи классификации методом ближайшего соседа справедливо следующее выражение функционала полного скользящего контроля Q_k:

$$Q_k(\mu) = \sum_{m=1}^{k} P(m)C(m), \text{ где } C(m) = C_{L-1-m}^{\ell-1}/C_{L-1}^{\ell}.$$

Теперь рассмотрим более сложный метод обучения μ_Ω, который запоминает не всю обучающую выборку, а лишь подмножество эталонных объектов $\Omega \subseteq X^L$. На стадии классификации используется тот же алгоритм ближайшего соседа, но теперь ближайшие соседи выбираются только из Ω. Обозначим этот алгоритм через γ_Ω.

Обозначим через $r_m^\Omega(x_i)$ ошибку, возникающую при замене известной классификации объекта x_i на ответ $y(x_{im})$ на m-м соседе, $r_m^\Omega(x_i) = I(x_i,y(x_{im}))$, где x_{im} — m-ый объект из множества эталонов Ω, если упорядочить их по возрастанию расстояний до объекта x_i. Обратим внимание, что если $x_i \in \Omega$, то $m = 1, \ldots, |\Omega| - 1$, а если $x_i \in X^L \backslash \Omega$, то $m = 1, \ldots, |\Omega|$.

Профилем Ω-компактности выборки X^L называется функция $P^\Omega(m)$, выражающая долю объектов выборки, для которых правильный ответ не совпадает с правильным ответом на m-ом соседе из множества эталонов:

$$P^\Omega(m) = \frac{1}{L}\sum_{i=1}^{L} r_m^\Omega(x_i).$$

Теорема 2.1.2. Для задачи классификации методом ближайшего соседа справедливо следующее выражение функционала $Q_k(\mu_\Omega)$:

$$Q_k(\mu_\Omega) = \sum_{m=1}^{k} P^\Omega(m)C(m).$$

В работе рассматривается еще один возможный функционал, основанный на принципе полного скользящего контроля, для алгоритма ближайшего соседа с эталонным множеством Ω, и доказывается, что он сводится к частоте ошибок $\nu(\gamma_\Omega, X^L \backslash \Omega)$.

Таким образом функционал $Q_k(\mu_\Omega)$ и частоту ошибок $\nu(\gamma_\Omega, X^L \backslash \Omega)$ тоже можно рассматривать как оценку компактности выборки X^L. В отличии

от Q_k значения $Q_k(\mu_\Omega)$ и $\nu(\gamma_\Omega, X_G^L \backslash \Omega)$ зависят не только от компактности выборки, но и от того, насколько хорошо множество Ω описывает структуру классов. Отсюда возникает ряд оптимизационных постановок задачи отбора объектов. Оценки вычислительной сложности получаемых постановок сформулированы в виде ряда теорем и утверждений.

Теорема 2.2.1. Задача поиска минимального размера вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов Ω минимальной мощности, по которому классификация алгоритмом ближайшего соседа γ_Ω даст $\nu(\gamma_\Omega, X_G^L \backslash \Omega) = 0$. Причем выборка X_G^L строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.

Следствие 2.2.0.1. Задача о поиске эталонного множества Ω размера, не превышающего h и дающего $\nu(\gamma_\Omega, X^L \backslash \Omega) = 0$, является NP-трудной.

Замечание: задача поиска множества эталонных объектов Ω, без ограничения по размеру, и дающего $\nu(\gamma_\Omega, X^L) = 0$, не является NP-трудной, поскольку имеет тривиальное решение $\Omega = X^L$.

Теорема 2.2.2. Задача поиска минимального размера вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов Ω, минимизирующего функционал $Q_1(\mu_\Omega)$. Причем выборка X_G^L строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$. Предполагается, что в Ω входит хотя бы по одному объекту из каждого класса.

Теорема 2.2.3. Задача поиска минимального вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов Ω минимальной мощности, минимизирующего функционал $Q_k(\mu_\Omega)$. Причем выборка X_G^L строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$. Предполагается, что в Ω содержится не менее $k+1$ объектов каждого класса.

Теорема 2.2.4. В условии теоремы 2.2.3 предположение о том, что в Ω лежит не менее $k + 1$ объектов каждого класса, можно заменить предположением $|\Omega| \geq k + 1$.

Следствие 2.2.0.2. Для любого k задача поиска множества эталонных объектов Ω, $|\Omega| \geq k+1$, для которого функционал $Q_k(\mu_\Omega)$ не превышает h, является NP-трудной.

Замечание: условие $|\Omega| \geq k+1$ действительно необходимо, иначе функционал Q невозможно будет записать в виде, указанном в теореме 2.1.2.

NP-трудность этих задач обосновывает применеие различных эвристических алгоритмов, выбирающих Ω так, что минимизируемый функционал принимает значение не наименьшее, но близкое к наименьшему, и/или мощность множества Ω не минимальна, но близка к минимальной.

Далее рассматривается задача отбора признаков. Пусть дана обучающая выборка X^L, описываемая множеством признаков $\mathbb{F}$. Компактность выборки зависит от функции расстояния ρ, при помощи которой измеряется сходство объектов. Для оценки компактности выборки будем использовать первый профиль компактности $P(1)$, который совпадает со значением функционала Q_1 для алгоритма ближайшего соседа.

Функция ρ обычно строится по признаковым описаниям объектов. Одними из самых используемых функций ρ является взвешенное расстояние Минковского:

$$\rho_{p\text{Mink}}(x, x') = \Big(\sum_{d=1}^{n} w_d \rho_d^p(x,x')\Big)^{\frac{1}{p}},$$

где $\rho_d(x,x')$ — расстояние между объектами x и x' по признаку f_d, а w_d — вес признака f_d, $w_d \geq 0$, $\sum_{d=1}^{|\mathbb{F}|} w_d = 1$.

Оценку компактности выборки с функцией расстояний $\rho_{p\text{Mink}}$ при помощи функции $P(1)$ можно свести к следующему частному случаю: признаки $f \in \mathbb{F}$ заданы матрицей попарных расстояний $\rho_f(x,x')$ между объектами, а возможные функции расстояний $\rho_{\mathbb{F}}(x,x')$ — линейными комбинациями $\rho_f(x,x')$, $f \in \mathbb{F}$ с неотрицательными весами. Для сведения задачи с $\rho_{p\text{Mink}}$ к задаче с $\rho_{\mathbb{F}}$ достаточно каждый элемент матриц попарных расстояний возвести в степень p.

Минимальное значение профиля компактности $P(1)$, достижимое при использовании функций расстояний, построенных на множестве признаков F, будем называть минимумом $P(1)$ на F.

Любая функция расстояний ρ_F, построенная на подмножестве признаков $F \subseteq \mathbb{F}$, может быть построениа и на $\mathbb{F}$. Следовательно, множество

признаков $\mathbb{F}$ всегда не хуже любого своего подмножества $F \subseteq \mathbb{F}$ с точки зрения $P(1)$. Однако чем больше признаков мы используем, тем сильнее модель склонна к переобучению. Задача отбора признаков ставится в данной главе следующим образом:

Задача 2.3.0.1. Выбрать минимальное по количеству множество признаков $F \subseteq \mathbb{F}$ такое, что минимум $P(1)$ на $\mathbb{F}$ совпадает с минимумом $P(1)$ на F.

Теорема 2.3.1. Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 2.3.0.1 на некоторой искусственной обучающей выборке X^L.

Следствие 2.3.0.1. Задача выбора множества признаков $F \subseteq \mathbb{F}$ такого, что минимум $P(1)$ на $\mathbb{F}$ совпадает с минимумом $P(1)$ на F и $|F| \leq q$ является NP-трудной.

Третья глава посвящена задачам отбор эталонов и признаков с ограничениями монотонности.

Задачи отбора объектов и признаков являются важным этапом построения монотонных классификаторов. Монотонными называются классификаторы, для которых предполагается наличие монотонных зависимостей между значениями признаков и меткой классов.

Многие методы монотонной классификации требуют монотонности обучающей выборки. Однако на практике монотонность на некоторых объектах может нарушаться. Потому приходится производить *монотонизацию* обучающей выборки — искать подмножества объектов и признаков, для которых будут выполнятся ограничения монотонности.

В данной главе устанавливается связь между метрическими и монотонными классификаторами. Использование монотонной функции расстояния, предложенной в работах К. В. Воронцова и Г. А. Махиной позволяет строить монотонный метрический классификатор. В данном разделе предложена более простая конструкция монотонной функции расстояния, которая позволяет оценить точность работы монотонных классификаторов произвольного вида.

Дана обучающая выборка $X^L = (x_i, y_i)_{i=1}^L$ и конечное множество признаков $\mathbb{F} = \{f_1, \ldots, f_t\}$ — отображений вида $f_j \colon X^L \to E_j$, где E_j — линейно упорядоченное множество. Каждый объект x_i относится к одному из двух

классов $y_i = y^*(x_i) \in \{0, 1\}$. Обозначим множество объектов обучающей выборки класса 1 через $\mathbb{A}$, а класса 0 — через $\mathbb{B}$.

Любое непустое подмножество множества признаков $F \subseteq \mathbb{F}$ индуцирует отношение частичного порядка на X^L: $x \leq x'$ тогда и только тогда, когда $f(x) \leq f(x')$ для всех $f \in F$; $x < x'$ тогда и только тогда, когда $x \leq x'$ и $x \neq x'$.

Пара объектов $(a,b) \in \mathbb{A} \times \mathbb{B}$ называется монотонной, если $a > b$. Множество всех монотонных пар обозначается через M.

Пара объектов $(a,b) \in \mathbb{A} \times \mathbb{B}$ называется дефектной, если $a < b$. Множество всех дефектных пар обозначается через D.

Множество пар, монотонных по признаку f, будем обозначать через M_f, монотонных по совокупности признаков F — через M_F. Множество пар, дефектных по признаку f, будем обозначать через D_f, дефектных по совокупности признаков F — через D_F.

Выборка называется *монотонной*, если её объекты не образуют ни одной дефектной пары.

Классификатор γ называется *монотонным*, если для любых двух объектов x, x' из $x < x'$ следует $\gamma(x) \leq \gamma(x')$.

Пусть дано некоторое монотонное эталонное множество Ω объектов с метками классов, которые монотонный алгоритм γ классифицирует правильно. Тогда любые объекты, большие какого-то объекта класса 1 из множества Ω, а также объекты, меньшие какого-то объекта класса 0 из множества Ω, будут однозначно классифицироваться исходя из свойств монотонности. Классификация остальных объектов будет неоднозначной и зависеть от конструкции алгоритма γ.

Если же множество эталонов Ω немонотонно, то монотонный алгоритм γ, обученный на множестве Ω, не сможет классифицировать все объекты из Ω в соответствии с известными метками классов. Кроме того, дефектные пары Ω могут провоцировать ошибки алгоритма γ на других объектах.

Объект u *доминирует* над объектом v если:

– объекты u и v принадлежат классу 1, и $v > u$

– объекты u и v принадлежат классу 0, и $v < u$

Объект v будем называть *недоминируемым*, если не существует объекта $u \in X^L$, который бы доминировал над объектом v.

Будем называть *пессимистичным монотонным классификатором с эталонным множеством* Ω и обозначать γ^p_Ω классификатор, который:

- правильно классифицирует объект u класса 1 тогда и только тогда, когда существует объект $a \in \mathbb{A} \cap \Omega$ такой, что $a \leq u$, и не существует объекта $b \in \mathbb{B} \cap \Omega$ такого, что $u \leq b$
- правильно классифицирует объект v класса 0 тогда и только тогда, когда существует объект $b \in \mathbb{B} \cap \Omega$ такой, что $v \leq b$, и не существует объекта $a \in \mathbb{A} \cap \Omega$ такого, что $a \leq v$

Обозначим через Ω_M подмножество объектов Ω, не участвующих в дефектных парах. Пессимистичный монотонный классификатор будет правильно классифицировать те и только те объекты, которые правильно классифицировал бы любой монотонный алгоритм, работающий корректно на множестве объектов Ω_M. То есть качество работы пессимистичного монотонного классификатора позволяет получить оценку снизу для качества работы монотонных классификаторов, корректно работающих Ω_M.

В данной главе показано, что пессимистический монотонный классификатор можно представить в виде классификатора ближайшего соседа со специальной функцией расстояния, которая является упрощенным вариантом монотонной функции расстояний, предложенной Г. А. Махиной.

Рассмотрим задачи отбора объектов и признаков с условием сохранения ограничения монотонности.

Задача 3.2.1.1. Выбрать признаки так, чтобы при отсутствии дефектных пар получить максимальное возможное при этом количество монотонных пар: $FS(|D| = 0 \colon |M| \to \max)$.

Теорема 3.2.1. Решение задачи о покрытии множества подмножествами сводится к решению задачи 3.2.1.1 на некоторой искусственной обучающей выборке X^L.

Следствие 3.2.1.1. Задача $FS(|D| = 0 \colon |M| \geq m)$ является NP-трудной.

Задача 3.2.1.2. Выбрать минимальное количество признаков таким образом, чтобы дефектные пары отсутствовали: $FS(|D| = 0 \colon |F| \to \min)$

Теорема 3.2.2. Решение задачи о покрытии множества подмножествами сводится к решению задачи 3.2.1.2 на некоторой искусственной обучающей выборке X^L.

Следствие 3.2.1.2. Задача $FS(|D| = 0 \colon |F| \leq q)$ является NP-трудной.

Задача 3.2.2.1. Выбрать из X^L минимальное множество эталонных объектов Ω такое, что пессимистичный монотонный классификатор с эталонным множеством Ω будет классифицировать все объекты X^L правильно, то есть частота ошибок будет равна нулю: $\nu(\gamma^p_\Omega, X^L) = 0$

Теорема 3.2.3. Минимальное множество эталонных объектов $\Omega \subseteq X^L$ такое, что $\nu(\gamma^p_\Omega, X^L) = 0$, является множеством всех недоминируемых объектов X^L.

Следствие 3.2.2.1. Задача поиска минимального множества эталонных объектов $\Omega \subseteq X^L$ такого, что $\nu(\gamma^p_\Omega, X^L) = 0$, является полиномиальной по количеству объектов в выборке.

Задача 3.2.2.2. Выбрать из X^L минимальное множество эталонных объектов Ω так, чтобы значение функционала $Q_k(\mu_\Omega)$ пессимистичного монотонного классификатора с эталонным множеством Ω было минимальным.

Теорема 3.2.4. Решение задачи о покрытии множества подмножествами сводится к решению задачи 3.2.2.2 на некоторой искусственной обучающей выборке X^L.

Следствие 3.2.2.2. Задача выбора из X^L множества эталонных объектов Ω, такого, что $|\Omega| \leq t$ и значение функционала $Q_1(\mu_\Omega)$ пессимистичного монотонного классификатора с эталонным множеством Ω минимально, является NP-трудной.

Далее рассматривается задача построения монотонного классификатора по немонотонной выборке.

Утверждение 3.3.0.1. Для произвольного подмножества признаков $F \subseteq \mathbb{F}$

$$M_F = \bigcap_{f \in F} M_f, \quad D_F = \bigcap_{f \in F} D_f.$$

Следствие 3.3.0.1. Для любых подмножеств признаков F, F'

$$F \subseteq F' \;\Rightarrow\; M_{F'} \subseteq M_F,\; D_{F'} \subseteq D_F \;\Rightarrow\; |M_{F'}| \leq |M_F|,\; |D_{F'}| \leq |D_F|. \quad (1)$$

Задача построения монотонного классификатора заключается в аппроксимации неизвестной функции $y^*\colon \mathbb{X} \to \{0,1\}$, заданной в точках обучающей выборки, монотонной функцией $y\colon \mathbb{X} \to \{0,1\}$. По определению монотонной функции для любых двух объектов x,$x' \in \mathbb{X}$ из $x < x'$ следует $y(x) \leq y(x')$. На практике обучающая выборка может не удовлетворять усло-

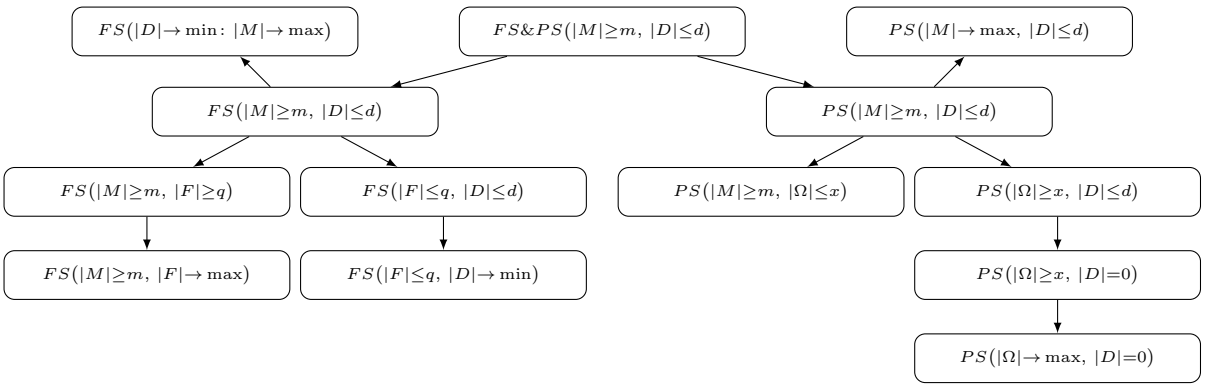

Рис. 1 — Постановки задачи монотонизации:
FS — отбор признаков, PS — отбор объектов.

вию монотонности, и в таких случаях ставится задача её предварительной монотонизации.

Задача монотонизации обучающей выборки состоит в том, чтобы выбрать подмножество объектов $\Omega \subseteq X^L$ и подмножество признаков $F \subseteq \mathbb{F}$ так, чтобы среди пар объектов (a,b) из $(\mathbb{A} \cap \Omega) \times (\mathbb{B} \cap \Omega)$ оказалось как можно меньше дефектных пар и как можно больше монотонных: $|D_F| \to \min$, $|M_F| \to \max$. Согласно следствию 3.3.0.1, первое условие можно заменить косвенным требованием $|F| \to \max$ или $|\Omega| \to \min$, а второе — косвенным требованием $|F| \to \min$ или $|\Omega| \to \max$.

Систематизация возникающих при этом постановок задач показана на рис. 1. Стрелками соединены те задачи, которые получаются друг из друга изменением одного из условий оптимальности. Вычислительная сложность задачи отбора объектов (PS) и отбора признаков (FS) рассматриваются отдельно.

Задача 3.3.1.1. Выбрать признаки так, чтобы при минимальном количестве дефектных пар получить максимальное возможное при этом количество монотонных пар: $FS(|D| \to \min\colon |M| \to \max)$.

Теорема 3.3.1. Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 3.3.1.1 на некоторой искусственной обучающей выборке X^L.

Задача 3.3.1.2. Выбрать множество признаков F так, чтобы монотонных пар было не менее m, а дефектных не более d: $FS(|D_F| \leq d, |M| \geq m)$.

Теорема 3.3.2. Задача $FS(|D_F| \leq d, |M| \geq m)$ является NP-трудной.

Задача 3.3.1.3. Получить не менее m монотонных пар, выбрав не менее q признаков: $FS(|M| \geq m, |F| \geq q)$.

Теорема 3.3.3. Задача $FS(|M| \geq m, |F| \geq q)$ является NP-трудной.

Задача 3.3.1.4. Получить минимальное количество дефектных пар, выбрав не более q признаков: $FS(|D_F| \to \min, |F| \leq q)$.

Теорема 3.3.4. Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 3.3.1.4 на некоторой искусственной обучающей выборке X^L.

Задача 3.3.1.5. Получить не более d дефектных пар, выбрав не более q признаков: $FS(|F| \leq q, |D_F| \leq d)$.

Теорема 3.3.5. Задача $FS(|F| \leq q, |D_F| \leq d)$ является NP-трудной.

Задача 3.3.2.1. Выбрать объекты так, чтобы монотонных пар было не менее m, а дефектных пар не более d: $PS(|D| \leq d, |M| \geq m)$.

Гипотеза. Задача $PS(|D| \leq d, |M| \geq m)$ является NP-трудной.

Доказательство данной гипотезы пока остаётся открытой проблемой.

Задача 3.3.2.2. Выбрать не более x объектов так, чтобы число монотонных пар было не менее m: $PS(|M| \geq m, |\Omega| \leq x)$.

Теорема 3.3.6. Задача $PS(|M| \geq m, |\Omega| \leq x)$ является NP-трудной.

Задача 3.3.2.3. Устранить все дефектные пары, оставив в обучающей выборке как можно больше объектов: $PS(|D| = 0, |\Omega| \to \max)$.

Задача 3.3.2.4. Устранить все дефектные пары, выбрав не менее x объектов обучающей выборки: $PS(|D| = 0, |\Omega| \geq x)$.

Обе задачи 3.3.2.3 и 3.3.2.4 имеют полиномиальное решение, время поиска удаляемых объектов составляет $O(d(a_d + b_d)^{0,5})$ по теореме Кёнига.

Задача 3.3.2.5. Получить не более d дефектных пар, выбрав не менее x объектов обучающей выборки X^L: $PS(|D| \leq d, |\Omega| \geq x)$.

Теорема 3.3.7. Задача 3.3.2.5 является полиномиальной по числу объектов.

Почти все предложенные формулировки задачи монотонизации являются NP-трудными. Из этого следует, что задачу отбора признаков и объектов для построения классификатора с ограничениями монотонности не всегда возможно и целесообразно решать точно, что обосновывает использование приближенных алгоритмов.

Четвертая глава посвящена построению алгоритма монотонизации с одновременным отбором объектов и признаков и его экспериментальной проверке. Если рассматривать задачи отбора объектов и признаков последовательно, возникает следующая проблема: неинформативные признаки мешают

отбирать объекты, делая почти все объекты обучающей выборки несравнимыми, а шумовые объекты могут не позволить удалить неинформативные признаки без нарушения условия монотонности. Одновременный отбор объектов и признаков предотвращает попадание в локальные минимумы, связанных с шумовыми объектами и неинформативными признаками.

В начале главы описываются общая схема жадного алгоритма монотонизации. Далее обсуждаются возможность применения известных функционалов, характеризующих монотонность выборки, в качестве целевой функции жадного алгоритма. Предлагается функционал, характеризующий монотонность выборки, адаптированный под идею одновременного отбора объектов и признаков. В целях экспериментальной проверки алгоритма монотонизации используются данные информационного анализа электрокардиосигналов для скрининговой диагностики, предоставленные д.м.н. проф. В. М. Успенским.

Для измерения качества классификации в медицинской диагностике принято использовать критерии чувствительности и специфичности. *Чувствительность* — это доля больных, для которых диагностическое правило верно диагностирует наличие болезни. *Специфичность* — это доля здоровых, для которых диагностическое правило верно диагностирует отсутствие болезни. В данной работе качество классификации оценивается с помощью критерия ROC-AUC (Area Under Curve) — площади под графиком зависимости чувствительности от специфичности. Преимущество критерия ROC-AUC в том, что он не зависит от выбора компромисса между чувствительностью и специфичностью. Его можно интерпретировать как долю правильно упорядоченных пар прецедентов.

Ход итерационного процесса отбора объектов и признаков (алгоритм *mon*) показан для одной из задач диагностики на графиках 2 и 3. График на рисунке 2 показывает зависимость числа монотонных и дефектных пар, числа признаков, а также числа объектов каждого из классов в обучающей подвыборке от номера итерации. На графике 3 изображены результаты следующего эксперимента: логистическая регрессия с регуляризатором L1 обучалась на обучающей подвыборке, получаемой на каждой итерации алгоритма *mon* (всего 249 классификаторов); для каждого полученного классификатора вычислялось значение ROC-AUC на контрольной подвыборке. Для сравнения на графике приведено значение ROC-AUC на контрольной подвыборке

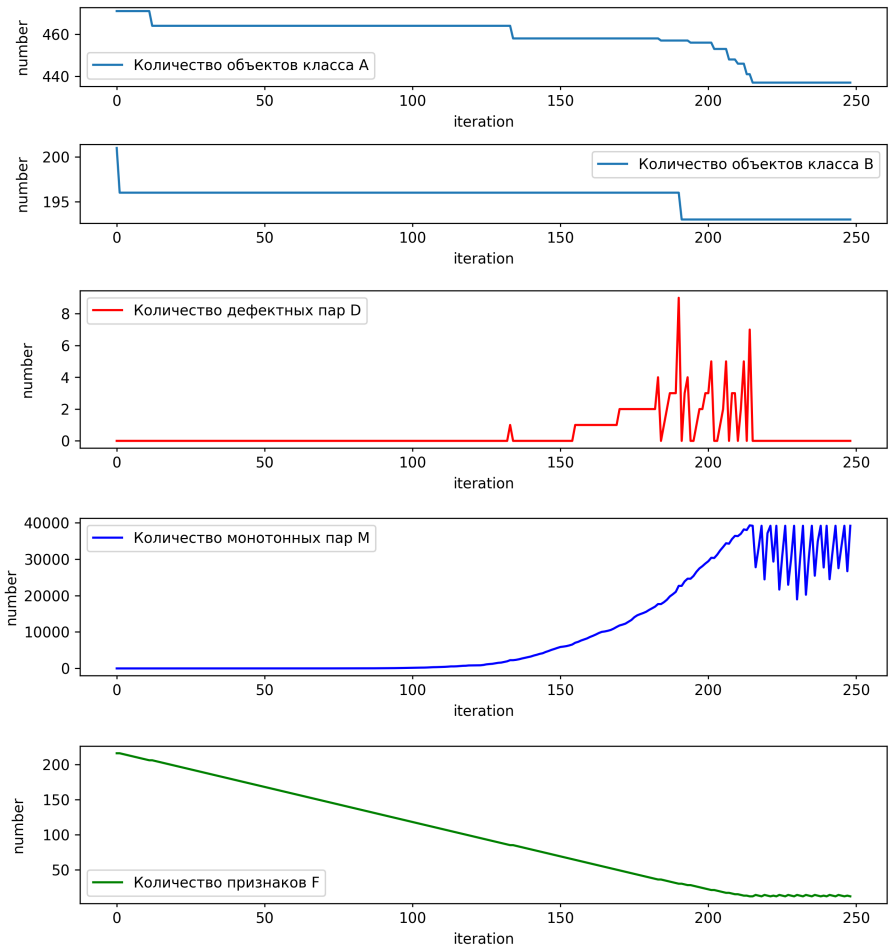

Рис. 2 — Зависимость числа объектов каждого из классов, дефектных пар, монотонных пар а также числа признаков в обучающей подвыборке от номера итерации алгоритма *mon*.

для логистической регрессии с регуляризатором L1, обученной на обучающей подвыборке без предварительной монотонизации.

В приложении к работе приводятся результаты экспериментов, показывающие, что для многих задач дифференциальной диагностики использование предложенного алгоритма монотонизации *mon* с одновременным отбором объектов и признаков существенно улучшает качество классификации. Результаты экспериментов иллюстрируются графиками и таблицами.

В **заключении** перечислены результаты, выносимые на защиту. В **приложении** приведены графики, иллюстрирующие работу предложенного алгоритма монотонизации с одновременным отбором объектов и признаков.

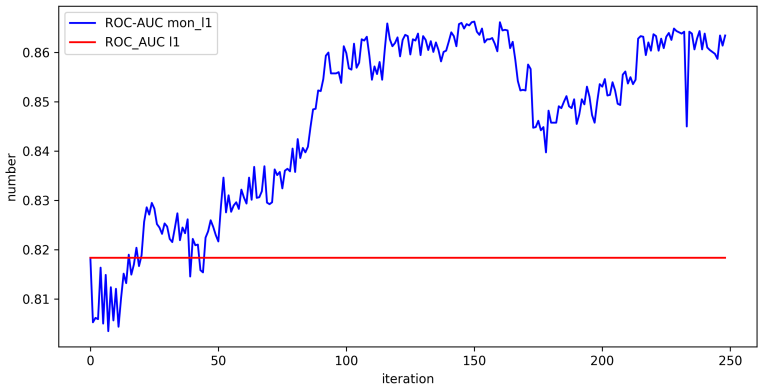

Рис. 3 — Значения ROC-AUC на контрольной подвыборке в зависимости от количества итераций алгоритма *mon* на обучающй подвыборке.

Основные результаты диссертации

1. Получены оценки вычислительной сложности задач отбора объектов и признаков для алгоритма ближайшего соседа. В предложенных постановках данные задачи являются NP-трудными.
2. Получены оценки вычислительной сложности задач отбора объектов и признаков для монотонизации обучающей выборки, предложена их систематизация. Почти все задачи являются NP-трудными, что обосновывает применение приближенных алгоритмов для их решения. Для единственной полиномиальной постановки указан точный эффективный алгоритм решения.
3. Предложен и протестирован экспериментально алгоритм монотонизации выборки с одновременным отбором объектов и признаков. Показано, что в ряде случаев использование предложенного алгоритма повышает качество классификации.

Публикации автора по теме диссертации

1. *Зухба А.В.* NP-полнота задачи оптимального отбора эталонных объектов в методе ближайшего соседа // Труды 52-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». Часть VII. Управление и прикладная математика. Том 2. — М: МФТИ, 2009. — С. 61–63.

2. *Зухба А.В.* Оценка оптимальности жадного алгоритма отбора эталонных объектов в методе ближайшего соседа // Труды 53-й научной конференции МФТИ «Совре-

менные проблемы фундаментальных и прикладных наук». Часть VII. Управление и прикладная математика. Том 2. — М: МФТИ, 2010. — С. 75–76.

3. *Зухба А.В.* Сложность задачи отбора эталонов в методе ближайшего соседа // Математические методы распознавания образов: 15-я Всероссийская конференция, г.Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М: МАКС Пресс, 2011. — С. 305–308.

4. *Зухба А.В.* Оценка вычислительной сложности задачи монотонизации выборки // Математические методы распознавания образов: 16-я Всероссийская конференция, г.Казань, 6–12 сентября 2013 г.: Тезисы докладов. — М: ТОРУС ПРЕСС, 2013. — С. 39.

5. *Зухба А.В.* Отбор объектов и признаков для монотонных классификаторов // Математические методы распознавания образов: Тезисы докладов 17-й Всероссийской конференции с международным участием, г.Светлогорск, 2015 г. — М: ТОРУС ПРЕСС, 2015. — С. 96–97.

6. *Швец М.Ю. Зухба А.В. Воронцов К.В.* Построение монотонного классификатора для задач медицинской диагностики // Математические методы распознавания образов: Тезисы докладов 17-й Всероссийской конференции с международным участием, г.Светлогорск, 2015 г. — М: ТОРУС ПРЕСС, 2015. — С. 42–43.

7. *Зухба А.В.* Алгоритм монотонизации выборки с одновременным отбором объектов и признаков // Математические методы распознавания образов: Тезисы докладов 18-й Всероссийской конференции с международным участием, г.Таганрог, 2017 г. — М: ТОРУС ПРЕСС, 2017. — С. 52.

8. *Zukhba A. V.* NP-Completeness of the Problem of Prototype Selection in the Nearest Neighbor Method // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, №.4. – P. 484–494.

9. *Зухба А.В.* Вычислительная сложность отбора объектов и признаков для задач классификации с ограничениями монотонности [Электронный ресурс] // *Математическая биология и биоинформатика.* — 2015. — Т. 10, № 2. — С. 356–371. — Режим доступа: http://www.matbio.org/article.php?journ_id=22&id=244. – (Дата обращения: 25.01.2018).

10. *Зухба А.В.* Оценка вычислительной сложности задачи монотонизации множества при помощи отбора признаков // *Математические и информационные модели управления: сб. науч. трудов.* — М: МФТИ, 2013. — С. 124–132.

Зухба Анастасия Викторовна

ОЦЕНКА ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ ЗАДАЧ ОТБОРА ЭТАЛОННЫХ ОБЪЕКТОВ И ПРИЗНАКОВ

Автореферат

Подписано в печать 22.02.2018. Формат 60x84 1/16. Усл. печ. л. 1,0.
Тираж 100 экз. Заказ № 375.

Федеральное государственное автономное образовательное учреждение высшего образования
«Московский физико-технический институт (государственный университет)»
Отдел оперативной полиграфии «Физтех-полиграф»
141700, Московская обл., г.Долгопрудный, Институтский пер., 9.