На правах рукописи

Гончаров Юрий Владимирович

Минимаксный подход к построению оптимального классификатора методом SVM с одновременным выбором оптимального подпространства признаков

01.01.09 — Дискретная математика и математическая кибернетика

Автореферат

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва — 2010

Работа выполнена в Учреждении Российской академии наук
Вычислительный центр им. А. А. Дородницына РАН

Научный руководитель:	доктор физико-математических наук, профессор А. С. Антипин
Официальные оппоненты:	доктор физико-математических наук А. М. Райгородский
	кандидат физико-математических наук А. Я. Червоненкис
Ведущая организация:	Учреждение Российской академии наук Институт проблем передачи информации им. А. А. Харкевича РАН

Защита состоится 10 июня 2010 г. в 13:00 часов на заседании диссертационного совета Д.002.17.02 при Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, г.Москва, ул.Вавилова, 40, конференц-зал.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук Вычислительный центр им. А. А. Дородницына РАН.

Автореферат разослан «____» _____ 2010 г.

Ученый секретарь диссертационного совета	доктор физико-математических наук, профессор В. В. Рязанов

Общая характеристика работы

В диссертационной работе рассматривается проблема выбора признаков в задаче обучения классификации. Предлагается минимаксный подход к одновременному построению оптимального решающего правила и оптимального подпространства признаков для классификации. Подход применяется к задаче метода опорных векторов (support vector machine, SVM) и приводит к минимаксной задаче оптимизации модифицированного критерия задачи SVM. Устанавливаются математические свойства решений минимаксной задачи. Предлагаются алгоритмы решения минимаксной задачи. Описываются численные эксперименты по тестированию предложенного подхода в задачах классификации. Демонстрируется применение предложенного минимаксного подхода к задаче одновременного построения SVM регрессии и оптимального подпространства признаков.

Актуальность темы. Определение множества информативных признаков рассматривалось в качестве одной из главных задач с самого начала изучения проблемы обучения классификации. Так, в одной из самых первых работ по распознаванию образов[1] построение множества «полезных» признаков рассматривалось в качестве необходимой компоненты в алгоритмах обучения классификации. С прикладной точки зрения необходимость выбора признаков диктуется тем, что обучающие данные часто содержат избыточные или не имеющие отношения к изучаемым явлениям признаки, которые могут значительно снизить качество распознавания. Наряду с задачей выбора признаков также выделяют задачу извлечения признаков, в которой из существующих признаков могут конструироваться новые признаки. При постановке задачи обучения объекты распознавания представляются векторами пространства $\mathbf{R}^n$, где n — количество признаков, характеризующих объект. Таким образом, при решении задачи выбора или извлечения признаков стремят-

[1] Бонгард М.М. Проблема узнавания. М.: Наука, 1967.

ся сократить размерность пространства. Предлагаемый в диссертации метод относится к группе «встроенных» методов, в которых признаки выбираются одновременно с процессом решения задачи обучения классификации. «Встроенные» методы представляют наибольший интерес в виду того, что эти методы позволяют учитывать особенности алгоритма обучения в процессе поиска подмножества информативных признаков. Среди огромного разнообразия современных подходов к распознаванию метод SVM на протяжении последних 15 лет общепризнан как один из самых совершенных. Следовательно, разработка «встроенных» методов выбора признаков на основе SVM представляется актуальной задачей. Поскольку в методе опорных векторов формулируется и точно решается оптимизационная задача в пространстве исходных признаков, то естественно попытаться найти обобщение имеющейся формулировки, которая бы обеспечивала поиск оптимального правила классификации при рассмотрении всех возможных подмножеств признаков. Существует несколько примеров таких обобщений задачи SVM. Так, в работе В.Н. Вапника и др.[2] предложена постановка задачи выбора признаков, метод решения которой сводится к минимизации дифференцируемой невыпуклой функции при ограничениях. В другой работе рассматривается постановка в форме задачи многокритериальной оптимизации[3], для которой предлагается приближенный алгоритм ее решения. В обеих постановках булевые переменные, которые отвечают за выбор подмножества признаков, заменяются на непрерывные переменные z_i со значениями из отрезка $[0, 1]$.

Проблемная ситуация заключается в том, что задачи оптимизации, возникающие в указанных постановках, решаются приближенными методами и не гарантируют нахождения даже локальных минимумов экстремизируемых функционалов.

[2]Weston, J., Mukherjee S., Chapelle O., Pontil M., Poggio T.,Vapnik V. Feature Selection for SVMs // Advances in Neural Information Processing Systems. 2000. V.13. P.668-674.

[3]Bi J. Multi-objective programming in SVMs // Proc. of 20th Intern.Conf. on Machine Learning. 2003. P.35-42.

Другая проблемная ситуация состоит в том, что значения переменных z_i, отличных от нуля и единицы в решении задач оптимизации, затрудняют их интерпретацию в качестве удаленных или выбранных признаков. В обеих постановках трудно судить об условиях, при которых переменные z_i в решении принимают целые значения.

В диссертации представлен подход к разрешению указанных проблемных ситуаций. Задача выбора признаков ставится на основе модификации оптимизационного критерия метода SVM. В постановке используются непрерывные переменные $z_i \in [0, 1]$, которые отвечают за выбор признаков. Получаемая задача оптимизации имеет выпуклую целевую функцию. Анализируются свойства решений задачи оптимизации и описываются условия, при которых значения переменных z_i принимают значения 0 или 1.

Цель работы. Целью диссертационной работы является разработка и экспериментальная проверка нового подхода к задаче выбора признаков на основе метода SVM при построении линейных оптимальных решающих правил классификации.

Задачи исследования. Для достижения цели работы поставлены следующие задачи:

1. Анализ современного состояния проблемы выбора признаков.
2. Разработка новой формулировки задачи выбора признаков в рамках метода опорных векторов.
3. Анализ свойств решений поставленной задачи выбора признаков.
4. Разработка алгоритма решения возникающих задач оптимизации и анализ сходимости алгоритма.
5. Разработка схемы для экспериментального тестирования разработанного алгоритма.
6. Реализация алгоритма решения задачи выбора признаков на ЭВМ, проведение и анализ результатов экспериментов.

Методологическая и теоретическая основа исследований. Теоретическую основу диссертации составили работы отечественных и зарубежных авторов в теории выпуклого анализа, оптимизации, прикладной статистики и распознавания образов. В диссертации использовались методы квадратичного программирования, вычисления седловых точек выпукло-вогнутых функций и методы недифференцируемой оптимизации субградиентного типа.

Научная новизна. В диссертации предложена оригинальная формулировка задачи выбора признаков, которая построена на модификации критерия метода опорных векторов. Математически задача выбора признаков поставлена в виде оптимизационной задачи, в которой булевой переменной z_i соответствует наличие или отсутствие в подмножестве информативных признаков i-го признака. Остальные переменные в задаче оптимизации отвечают за поиск решающей функции распознавания. Задача дискретной, по переменным z_i, оптимизации погружается в непрерывную невыпуклую задачу, в которой переменные z_i принимают значения из интервала $[0, 1]$. Показано, что невыпуклая задача может быть заменена на эквивалентную задачу на поиск минимакса выпукло-вогнутой функции. Доказана теорема о выпуклости по переменным z_i минимизируемой функции. Главным научным результатом диссертации является теорема, характеризующая условия, при которых оптимальные значения переменных z_i в минимаксной задаче принимают значения 0 или 1. Показано, что среди множества решений минимаксной задачи существуют решения, являющиеся седловыми точками целевой функции минимаксной задачи. Для решения задачи на минимакс предложено использовать алгоритм поиска седловой точки выпукло-вогнутой функции. Для поиска седловой точки применен дискретный вариант экстраградиентного метода[4], для которого доказана сходимость и проведена оценка параметра, задающего величину шага в алгоритме. В экстраградиентном методе для вы-

[4]Antipin A.S. From optima to equilibria // Proceedings of Institute for Systems Analysis. «Dynamics of non-homogeneous systems». V.3. Moscow. 2000. P.35-64.

числения проекции вместо решения задачи квадратичного программирования предложено использовать алгоритм чередующихся проекций Дейкстры, что значительно ускоряет работу метода.

Теоретическая и практическая значимость. Предложенный минимаксный подход к выбору признаков в задаче классификации может быть применен к другим задачам обработки данных. В приложении **C** диссертации показано применение минимаксного подхода при выборе признаков в задаче построения SVM регрессии. Разработано математическое обеспечение, позволяющее решать задачи выбора признаков при обучении классификации и построении регрессии методом опорных векторов. Математическое обеспечение решения задач выбора признаков реализовано в виде независимых модулей и может быть использовано независимыми разработчиками.

Апробация работы. Основные результаты работы докладывались на 13-й Всероссийской конференции «Математические методы распознавания образов» в 2007г.

Публикации автора. Основные результаты диссертации опубликованы в работах [1–6]. См. в конце автореферата.

Структура и объем работы. Диссертация состоит из введения, пяти глав, заключения, списка литературы и трех приложений. Список литературы содержит 89 наименований. Диссертация изложена на 174 страницах, содержит 21 рисунок и восемь таблиц.

Содержание работы. *Во введении* обосновывается актуальность диссертационной работы, определяются цели и задачи исследования, описывается научная новизна и дано краткое содержание диссертации.

В первой главе приводится постановка задачи обучения классификации, описывается задача сокращения размерности и дается обзор нескольких классических методов извлечения признаков.

Задача обучения классификации состоит в поиске решающего правила, согласно которому любой вектор $x \in \mathbf{R}^n$ относится к одному из двух классов. Правило может задаваться в виде решающей функции f. При $f(x) = 1$ вектор x относится к первому классу, при $f(x) = -1$ — ко второму. Задано множество допустимых решающих функций $\{f_\alpha(x),\ \alpha \in \Omega\}$, где Ω — некоторое множество. Поиск оптимальной функции $f_\alpha(x)$ ведется по обучающей выборке конечной длины $(x_1, y_1), (x_2, y_2), \ldots (x_l, y_l)$, где $x_i \in \mathbf{R}^n, y_i \in \{1; -1\}, i = 1, 2, \ldots, l$. Элемент обучающей выборки состоит из вектора x_i и y_i — признака принадлежности одному из двух классов.

Во второй главе дан обзор современного состояния проблемы выбора признаков. Описываются основные постановки задач выбора признаков и алгоритмы их решения. Перечисляются различные виды оценок качества подмножества признаков. Подробно рассказывается о методах оценки вероятности ошибки классификации.

В третьей главе дается описание метода опорных векторов и двух алгоритмов выбора признаков, построенных на основе этого метода. Анализируются проблемы и недостатки описанных алгоритмов.

Алгоритм метода опорных векторов для задачи классификации двух линейно-разделимых классов известен уже более 35 лет и назван в книге В.Н. Вапника и А.Я. Червоненкиса[5] методом обобщенного портрета. Почти 20 лет спустя было предложено обобщение алгоритма метода опорных векторов для задачи классификации на случай линейно неразделимых классов[6]. В англоязычной литературе метод имеет название «SVM (support vector machine)». Изложение метода опорных векторов в диссертации следует книге В.Н. Вапника.[7] «Емкость» множества решающих функций $\{f_\alpha(x),\ \alpha \in \Omega\}$ характеризуется величиной VC-размерности, которую будем обозначать че-

[5]Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974.

[6]Cortes C., Vapnik V. Support Vector Networks // Machine Learning. 1995. V.20. №3. P.273-297.

[7]Vapnik V.N. The Nature of Statistical Learning Theory, Second Edition. New York: Springer, 2000.

рез h. По обучающей выборке вычисляется величина эмпирического риска функции $f_\alpha(x)$:

$$R_{emp}(\alpha) = \frac{1}{2l}\sum_{i=1}^{l}|y_i - f_\alpha(x_i)|, \alpha \in \Omega\ .$$

Оценка сверху для вероятности ошибки распознавания решающим правилом $f_\alpha(x)$ имеет вид монотонно возрастающей функции $g(h, R_{emp}(\alpha))$ по h и $R_{emp}(\alpha)$. При поиске оптимального решающего правила $f_\alpha(x)$ в методе опорных векторов стремятся минимизировать величину оценки $g(h, R_{emp}(\alpha))$. Для этого стараются минимизировать взаимно противоречивые VC-размерность и эмпирический риск.

Встречающиеся далее обозначения нормы вектора и матрицы подразумевают соответственно стандартную евклидову норму вектора и подчиненную ей матричную норму. Через $\langle x_1, x_2\rangle$ обозначим скалярное произведение векторов x_1 и x_2. Пусть дана гиперплоскость $\langle w^*, x\rangle + b^* = 0$, $\|w^*\| = 1$ и число $\Delta > 0$. Рассмотрим следующее правило:

$$f(x) = \begin{cases} 1 & \text{, если } \langle w^*, x\rangle + b^* \geq \Delta, \\ -1 & \text{, если } \langle w^*, x\rangle + b^* \leq -\Delta. \end{cases} \tag{1}$$

В случае выполнения неравенств $-\Delta < \langle w^*, x_i\rangle + b^* < \Delta$ считается, что решающее правило классифицирует элемент обучающей выборки x_i с ошибкой.

В.Н. Вапником и А.Я. Червоненкисом получена верхняя оценка величины VC-размерности множества решающих правил (1):

$$h \leq min\left(\left\lceil\frac{R^2}{\Delta^2}\right\rceil, n\right) + 1, \tag{2}$$

где R — радиус шара, в котором лежат векторы x; $\lceil a\rceil$ — наименьшее целое число большее или равное a. Для минимизации величины h предлагается минимизировать правую часть оценки (2).

Метод опорных векторов сводится к решению задачи квадратичного про-

граммирования.

$$\min_{w,b,\delta_i} \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \delta_i \right) , \tag{3}$$
$$y_i(\langle w, x_i\rangle + b) \geq 1 - \delta_i, \quad \delta_i \geq 0, \quad C > 0, \quad i = 1, 2, \ldots, l \, .$$

Минимизация первого члена функционала (3) ведет к максимизации величины Δ в оценке (2). Минимизация второго члена функционала (3) ведет к минимизации эмпирического риска $R_{emp}(\alpha)$. В оценке (2) участвует величина размерности n, которую необходимо уменьшать для минимизации оценки VC-размерности. Однако, в функционале задачи опорных векторов (3) нет члена, который бы побуждал уменьшать величину размерности n. В диссертации предлагается критерий оптимизации, который штрафует также и величину размерности n.

В четвертой главе описывается оригинальная постановка задачи выбора признаков. Первоначально задача ставится в форме дискретной задачи оптимизации. Описывается преобразование дискретной задачи в минимаксную задачу оптимизации. Анализируются свойства решений минимаксной задачи и доказывается теорема об условиях «целочисленности» решений по переменным, отвечающим за выбор признаков. Приводится алгоритм поиска седловой точки для решения минимаксной задачи и доказывается его сходимость. Дается описание вычисления длины шага алгоритма при решении задачи выбора признаков. Описывается быстрое вычисления проекций для использования в алгоритме поиска седловой точки.

Дискретная постановка задачи выбора признаков.

Пусть $I = \{1; 2; \ldots; n-1; n\}$ — множество координат вектора $v \in \mathbf{R}^n$, $Q \subseteq I$ — подмножество координат. Обозначим через v^Q вектор с множеством координат Q, $v_i^Q = v_i$, $i \in Q$. Например, для $I = \{1; 2; 3; 4; 5\}, v = (9; 8; 4; 7; 6)^T, Q = \{2; 3; 5\}$, вектор v^Q будет равен $(8; 4; 6)^T$.

Задача выбора признаков ставится в форме модификации задачи SVM (3):

$$\min_{Q\subseteq I, w^Q, b, \delta} \left(\frac{1}{2}\left\|w^Q\right\|^2 + C\sum_{i=1}^{l}\delta_i + A\,|Q| \right), \tag{4}$$
$$y_i(\langle w^Q, x_i^Q\rangle + b) \geq 1 - \delta_i,$$
$$\delta_i \geq 0, \quad i = 1, 2, \ldots, l, \quad A > 0.$$

Первые два члена функционала в (4) отвечают за поиск оптимальной гиперплоскости в пространстве признаков, задаваемом Q. Третий член — шраф на мощность подмножества признаков. Коэффициент A регулирует величину штрафа. Задача (4) имеет комбинаторный характер по Q. Различным подмножествам Q отвечают различные подпространства признаков. Однако, вид целевой функции позволяет погрузить задачи с различными Q в исходное пространство признаков. Рассмотрим сначала следующую задачу:

$$\min_{z, w, b, \delta} \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\delta_i + A\sum_{j=1}^{n} z_j \right), \tag{5}$$
$$y_i \left(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b \right) \geq 1 - \delta_i, \tag{6}$$
$$\delta_i \geq 0, \quad i = 1, 2, \ldots, l, \quad A > 0, \quad z_j \in \{0; 1\}, \quad j = 1, 2, \ldots, n;$$

здесь x_i^j — координата j вектора x_i, w_j — координата j вектора w. Эта задача содержит булевые переменные z_j. Значение $z_j = 1$ означает, что признак с индексом j выбран, значение $z_j = 0$ означает, что признак удален.

Следующее предложение утверждает, что задачи (4) и (5),(6) эквивалентны в том смысле, что из решения одной задачи получаем решение другой.

Предложение 1 *Пусть Q, w^Q, b, δ — решение задачи (4), значения z, w в задаче (5),(6) вычисляются по следующему правилу:*

$$z_j = 1, w_j = w_j^Q \text{ , если } j \in Q,$$
$$z_j = 0, w_j = 0 \text{ , если } j \notin Q,$$

тогда z, w, b, δ *— решение задачи (5),(6).*

Пусть z, w, b, δ *— решение задачи (5),(6), значения* Q, w^Q *вычисляются по следующему правилу:*

$$Q = \{j | z_j = 1; j = 1, 2, \ldots, n\}, w_j^Q = w_j, j \in Q,$$

тогда Q, w^Q, b, δ *— решение задачи* (4).

Непрерывная постановка задачи выбора признаков. В задаче (5),(6) переменные z_j могут принимать значения 0 или 1. Задача оптимизации с целочисленными переменными является трудной для решения с вычислительной точки зрения. Предлагается ослабить условие на целочисленность и разрешить переменным z_j принимать значения из отрезка $[0, 1]$. Таким образом, приходим к непрерывному аналогу задачи (5),(6):

$$\min_{z,w,b,\delta} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \delta_i + A \sum_{j=1}^{n} z_j \right), \tag{7}$$

$$y_i(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b) \geq 1 - \delta_i, \tag{8}$$

$$\delta_i \geq 0, \quad i = 1, 2, \ldots, l, \quad A > 0, \quad z_j \in [0, 1], \quad j = 1, 2, \ldots, n\ .$$

Пусть z^*, w^*, b^*, δ^* — решение задачи (7),(8). Для классификации вектора x вычисляется следующая величина:

$$S = \sum_{j=1}^{n} w_j^* x^j \sqrt{z_j^*} + b^*.$$

При $S \geq 0$ вектор x относится к первому классу, в противном случае ко второму. Целевая функция задачи выбора признаков (7),(8) отличается от целевой функции SVM наличием штрафа $A \sum_{j=1}^{n} z_j$ на подмножество выбранных признаков. Присутствие в ограничениях (8) операции взятия квадратного корня $\sqrt{z_j}$ необходимо для того, чтобы целевая фунция двойственной к (7),(8) задачи была линейной по z. Задача (7),(8) при фиксированном z фактически

является задачей SVM, в которой элементы обучающей выборки предварительно шкалируются с помощью вектора z следующим способом:

$$x_i^j \longrightarrow x_i^j \sqrt{z_j}, \quad i = 1, 2, \ldots, l, \quad j = 1, 2, \ldots, n.$$

Можно считать, что процедура решения задачи (7),(8) состоит из поиска оптимальной шкалы признаков и оптимальной разделяющей гиперплоскости в новом шкалированном пространстве. Алгоритм, который решает задачу (7),(8), будем называть алгоритмом выбора признаков. Задача (7),(8) имеет невыпуклые ограничения по совокупности всех переменных z, w, b, δ.

Выпуклая минимаксная постановка задачи выбора признаков. Задача (7),(8) может быть преобразована в эквивалентную задачу последовательной минимизации с выпуклой структурой. Рассмотрим задачу:

$$\begin{aligned} &\min \psi(z), \\ &0 \le z_j \le 1, \quad j = 1, 2, \ldots, n, \end{aligned} \tag{9}$$

где значение $\psi(z)$ получается в результате решения следующей задачи:

$$\begin{aligned} &\psi(z) = \min_{w,b,\delta} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \delta_i + A \sum_{j=1}^{n} z_j \right), \\ &y_i \left(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b \right) \ge 1 - \delta_i, \\ &\delta_i \ge 0, \quad i = 1, 2, \ldots, l, \quad A > 0. \end{aligned} \tag{10}$$

Задача (10) может быть записана в двойственной форме

$$\psi(z) = \max_{\lambda} \left(\sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{k=1}^{l} \left(y_i y_k \sum_{j=1}^{n} z_j x_i^j x_k^j \right) \lambda_i \lambda_k + A \sum_{j=1}^{n} z_j \right), \tag{11}$$

$$\sum_{i=1}^{l} \lambda_i y_i = 0, \quad 0 \le \lambda_i \le C, \quad i = 1, \ldots, l.$$

Теорема 1 *Функция $\psi(z)$ в задаче (11) является выпуклой.*

Используя двойственное представление (11) для $\psi(z)$, задачу (9), (10) можно представить в форме минимаксной задачи (9), (11) и записать в следующем виде:

$$\min_{z\in Z}\max_{\lambda\in\Lambda} L(z,\lambda), \tag{12}$$

$$L(z,\lambda)=\sum_{i=1}^{l}\lambda_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}\left(y_iy_k\sum_{j=1}^{n}z_jx_i^jx_k^j\right)\lambda_i\lambda_k+A\sum_{j=1}^{n}z_j,$$

$$Z=\{z|0\le z_j\le 1, j=1,2,\ldots,n\}\ ,$$

$$\Lambda=\{\lambda|\sum_{i=1}^{l}\lambda_iy_i=0, 0\le\lambda_i\le C,\ i=1,2,\ldots,l\}\ .$$

Полученный результат о выпуклости функции $\psi(z)$ в задаче (11) позволяет построить алгоритм решения задачи (9),(11) одним из методов выпуклой недифференцируемой оптимизации. В приложении **В** диссертации описан алгоритм решения задачи (9),(11) на основе метода проекции субградиента Б.Т. Поляка.

Седловая постановка задачи выбора признаков. Заметим, что $L(z,\lambda)$ выпукло-вогнутая функция, т.е. выпуклая по z при фиксированном значении λ и вогнутая по λ при фиксированном z.

Рассмотрим задачу поиска седловой точки $(z^*,\lambda^*)\in Z\times\Lambda$:

$$L(z^*,\lambda)\le L(z^*,\lambda^*)\le L(z,\lambda^*)\quad\forall z\in Z,\forall\lambda\in\Lambda. \tag{13}$$

Для седловой точки справедливо следующее равенство:

$$\min_{z\in Z}\max_{\lambda\in\Lambda} L(z,\lambda)=\max_{\lambda\in\Lambda}\min_{z\in Z} L(z,\lambda)=L(z^*,\lambda^*). \tag{14}$$

Опираясь на это равенство, мы можем заменить задачу поиска минимакса на задачу поиска седловой точки. Множество седловых точек является подмножеством минимаксных решений.

Основным научным результатом диссертации служит следующая теорема «целочисленности», которая утверждает существование седловой точки у функции $L(z,\lambda)$ и показывает условия, при которых координаты оптимального z имеют целочисленные значения 0 или 1.

Теорема 2 *1. Существует седловая точка* (z^*,λ^*) *в задаче* (13) *и справедливы следующие равенства:*

$$z^* = \arg\min_{z\in Z}\max_{\lambda\in\Lambda} L(z,\lambda), \quad \lambda^* = \arg\max_{\lambda\in\Lambda}\min_{z\in Z} L(z,\lambda).$$

2. Если в седловой точке выполнено неравенство

$$\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* > 2A \ \textbf{, то } z_j^* = 1.$$

Если в седловой точке выполнено неравенство

$$\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* < 2A \ \textbf{, то } z_j^* = 0.$$

Если $0 < z_j^* < 1$, ***то*** *имеет место равенство*

$$\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* = 2A\ .$$

Может показаться, что в случае равенства

$$\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* = 2A\ ,$$

можно всегда избавиться от него малым варьированием параметра A. В диссертации приведен пример, когда при малом варьировании параметра A, равенство сохраняется.

Справедлива аналогичная теорема о «целочисленности» решений минимаксной задачи (12).

Теорема 3 *Пусть* (z^0,λ^0) *— решение минимаксной задачи* (12), (z^*,λ^*) *— седловая точка задачи* (13), *тогда справедливы следующие импликации.*

$$\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* > 2A \Rightarrow z_j^0 = 1,$$

$$\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* < 2A \Rightarrow z_j^0 = 0,$$

$$0 < z_j^0 < 1 \Rightarrow \sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* = 2A\ .$$

В диссертации анализируется влияние параметра A в задаче (12) на множество удаленных признаков. Следующее предложение показывает, что устанавливая достаточно большое значение параметра A в задаче (12), можно получить решение (z^0, λ^0), в котором для произвольного j будет выполнено равенство $z_j^0 = 0$.

Предложение 2 *Пусть величина A в задаче* (12) *вычислена по формуле*

$$A = \frac{1}{2}\max_{\lambda}\sum_{i=1}^{l}\sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i \lambda_k + \varepsilon,$$

$$\lambda \in \Lambda \quad , \varepsilon > 0,$$

где ε — произвольное положительное число и (z^0, λ^0) — решение задачи (12). *Тогда выполняется равенство $z_j^0 = 0$.*

С другой стороны, в диссертации приведен пример, в котором уменьшение параметра A в задаче (12) не приводит к решению с положительным значением j-той координаты вектора z.

Алгоритм поиска седловой точки. В этом разделе диссертации описывается алгоритм поиска седловой точки для выпукло-вогнутой фунции, которая удовлетворяет некоторым условиям на функцию $L(z, \lambda)$. Затем показывается, что функция $L(z, \lambda)$ в задаче (13) удовлетворяет этим условиям.

Пусть выпукло-вогнутая функция $L(z, \lambda)$ определена на декартовом произведении выпуклых замкнутых множеств $Z \times \Lambda$ и для некоторых констант $M_1 > 0$, $M_2 > 0$, $M_3 > 0$ удовлетворяет следующим неравенствам:

$$\left| L(z, \lambda + h) - L(z, \lambda) - \left\langle \frac{\partial L}{\partial \lambda}(z, \lambda), h \right\rangle \right| \leq \frac{1}{2} M_1 \left\| h \right\|^2, \tag{15}$$

$$\left| L(z+h,\lambda) - L(z,\lambda) - \left\langle \frac{\partial L}{\partial z}(z,\lambda), h \right\rangle \right| \le \frac{1}{2} M_2 \|h\|^2, \tag{16}$$

$$\left\| \frac{\partial L}{\partial z}(z,\lambda+h) - \frac{\partial L}{\partial z}(z,\lambda) \right\| \le M_3 \|h\|. \tag{17}$$

Пусть π_Z,π_Λ — операторы проекции на множества Z и Λ, т.е. $\pi_Z(z)$ — это проекция точки z на множество Z и, аналогично, $\pi_\Lambda(\lambda)$ — проекция точки λ на множество Λ.

Рассмотрим следующий алгоритм, каждая итерация которого состоит из следующих 3 шагов:

$$\begin{aligned} \bar{z}^k &= \pi_Z\left(z^k - \alpha \frac{\partial L}{\partial z}\left(z^k, \lambda^k\right)\right), \\ \lambda^{k+1} &= \pi_\Lambda\left(\lambda^k + \alpha \frac{\partial L}{\partial \lambda}\left(\bar{z}^k, \lambda^k\right)\right), \\ z^{k+1} &= \pi_Z\left(z^k - \alpha \frac{\partial L}{\partial z}\left(z^k, \lambda^{k+1}\right)\right). \end{aligned} \tag{18}$$

В диссертации доказывается теорема о сходимости алгоритма (18).

Теорема 4 *Пусть $L(z,\lambda)$ — выпукло-вогнутая функция на $Z \times \Lambda$, множества Z и Λ - выпуклые, замкнутые, $L(z,\lambda)$ удовлетворяет неравенствам (15)–(17), выполнены неравенства $0 < \alpha < \min\left(\frac{1}{M_2}, \frac{-M_1+\sqrt{M_1^2+8M_3^2}}{4M_3^2}\right)$. Тогда для любой начальной точки $z^0 \in Z$, $\lambda^0 \in \Lambda$ последовательность $(z^k,\lambda^k), k = 1,2,\ldots$, вычисляемая по формулам (18), сходится к (z^*,λ^*) — седловой точке функции $L(z,\lambda)$.*

Теорема 4 может быть распространена на случай, когда неравенство (16) выполняется при $M_2 = 0$. В этом случае в качестве константы Липшица в неравенстве (16) можно также взять $M_2 = \varepsilon$, где $\varepsilon > 0$ — произвольно малая величина. Для достаточно малой величины ε условие на величину параметра шага α в теореме 4 принимает вид $0 < \alpha < \frac{-M_1+\sqrt{M_1^2+8M_3^2}}{4M_3^2}$. Будем пользоваться последним условием на величину α при $M_2 = 0$.

Пусть определены матрицы

$$G = \{g_{ij} = y_i x_i^j | i = 1, 2, \ldots, l, j = 1, 2, \ldots, n\}, \tag{19}$$

$$R^j = \{R^j_{ik} = y_i y_k x_i^j x_k^j | i, k = 1, 2, \ldots, l\}, j = 1, 2, \ldots, n, \tag{20}$$

тогда константы M_1, M_2 и M_3 в (15)–(17) для $L(z, \lambda)$ из (13) вычисляются по следующим формулам:

$$M_1 = \|G\|^2, \quad M_2 = 0, \quad M_3 = Cl \sum_{j=1}^{n} \left\|R^j\right\|.$$

Быстрое вычисление проекций. В алгоритме (18) требуется вычислять проекции точек на множества Z и Λ. Обычно, если вычисляется проекция точки x_0 на множество S, задаваемое системой линейных неравенств, то решается задача квадратичного программирования:

$$\min \|x - x_0\|^2, x \in S. \tag{21}$$

В нашем случае возможно находить проекции более быстрыми с вычислительной точки зрения способами. Проекция точки $\widehat{z}$ на множество Z имеет вид

$$z^{pr} = \pi_Z(\widehat{z}), \quad Z = \{0 \le z_j \le 1, j = 1, 2, \ldots, n\}, \widehat{z} \in \mathbf{R}^n$$

и вычисляется по формуле

$$z_j^{pr} = \begin{cases} 0, & \widehat{z}_j < 0, \\ \widehat{z}_j, & 0 \le \widehat{z}_j \le 1, \\ 1, & \widehat{z}_j > 1. \end{cases} \quad j = 1, 2, \ldots, n, \tag{22}$$

Проекция точки $\widehat{\lambda}$ на множество Λ имеет вид

$$\lambda^{pr} = \pi_\Lambda(\widehat{\lambda}), \quad \Lambda = \left\{ \sum_{j=1}^{l} \lambda_j y_j = 0, 0 \le \lambda_j \le C, j = 1, 2, \ldots, l \right\}, \quad \widehat{\lambda} \in \mathbf{R}^l$$

и вычисляется приближенно в результате последовательных проекций на куб $\{0 \le \lambda_j \le C, j = 1, \ldots, l\}$ и на подпространство, задаваемое равенством $\sum_{j=1}^{l} \lambda_j y_j = 0$.

Проекция точки $\widehat{\lambda}$ на куб $\{0 \le \lambda_j \le C, j = 1, 2, \ldots, l\}$ вычисляется по формуле

$$\lambda_j^{pr} = \begin{cases} 0, & \widehat{\lambda}_j < 0, \\ \widehat{\lambda}_j, & 0 \le \widehat{\lambda}_j \le C, \quad j = 1, 2, \ldots, l, \\ C, & \widehat{\lambda}_j > C. \end{cases} \tag{23}$$

Проекция точки $x_0 \in \mathbf{R}^M$ на гиперплоскость $c^T x = 0, c \in R^M$ вычисляется по следующей формуле:

$$x_0^{pr} = x_0 - \frac{c^T x_0}{c^T c} c. \tag{24}$$

Таким образом, имеем формулы для вычисления проекций отдельно на куб и подпространство. Необходимо вычислять проекцию на их пересечение. Существует целый класс алгоритмов поиска проекции точки на пересечение множеств, использующий операции проекции отдельно на каждое множество. Одним из первых алгоритмов этого класса является алгоритм чередующихся проекций Дейкстры.

Пусть A, B — выпуклые замкнутые подмножества пространства $\mathbf{R}^l$ и $x \in \mathbf{R}^l$. Пусть для $k \ge 1$ определены последовательности

$$\begin{aligned} b_0 &= x, & p_0 &= q_0 = 0, \\ a_k &= \pi_A(b_{k-1} + p_{k-1}), & p_k &= b_{k-1} + p_{k-1} - a_k, \\ b_k &= \pi_B(a_k + q_{k-1}), & q_k &= a_k + q_{k-1} - b_k. \end{aligned} \tag{25}$$

Последовательности a_k, b_k сходятся к $\pi_{A \cap B}(x)$ — проекции точки x на пересечение множеств $A \cap B$.[8] В том случае, когда множество A является аффинным подпространством (сдвигом подпространства на вектор), то в формулах (25)

[8] Bauschke H.H., Borwein J.M. Dykstra's Alternating Projection Algorithm for Two Sets // J. Approximat. Theory. 1994. V.79. №3. P.418-443.

можно положить $p_k = 0$ для всех k.[9] Таким образом, алгоритм нахождения проекции точки x на множество $\Lambda = \{\sum_{j=1}^{l} \lambda_j y_j = 0, 0 \le \lambda_j \le C, j = 1, \ldots, l\}$ имеет следующий вид:

$$\begin{aligned} b_0 &= x, & q_0 &= 0, \\ a_k &= \pi_A(b_{k-1}), & & \\ b_k &= \pi_B(a_k + q_{k-1}), & q_k &= a_k + q_{k-1} - b_k, \end{aligned} \tag{26}$$

где множества A и B задаются в виде $A = \{\sum_{j=1}^{l} \lambda_j y_j = 0\}$, $B = \{0 \le \lambda_j \le C, j = 1, \ldots, l\}$ и проекции на множества A и B вычисляются по формулам (24) и (23).

В пятой главе диссертации описаны результаты численных экспериментов с разработанным алгоритмом выбора признаков. Эксперименты проводились на искусственных данных, данных для распознавания гласных звуков английского языка и медицинских данных для диагностики болезни. Результаты экспериментов показывают, что алгоритм способен одновременно удалять признаки и улучшать качество распознавания по сравнению с качеством распознавания SVM. В некоторых экспериментах алгоритм позволяет удалять признаки ценой незначительного ухудшения качества распознавания по сравнению с SVM. Проведены эксперименты по анализу вычислительной эффективности алгоритма вычисления проекции Дейкстры по сравнению с алгоритмом квадратичного программирования для использования в алгоритме выбора признаков. Общее время работы алгоритма выбора признаков уменьшается в 1.2 — 2.2 раза за счет использования алгоритма Дейкстры.

В приложении A дается доказательство предложения 1 (см. с.11) об эквивалентности двух постановок задачи выбора признаков.

В приложении B описан алгоритм решения минимаксной задачи (12) методом недифференцируемой оптимизации субградиентного типа.

В приложении C демонстрируется применение разработанного подхода

[9] Gaffke N., Mathar R. A cyclic projection algorithm via duality // Metrika. 1989. V.36. №1. P.29-54.

к выбору признаков в задаче построения SVM регрессии. Описываются постановка и математические свойства задачи, алгоритм решения и результат численного эксперимента с данными анализа ингибиторных свойств производных триазина от параметров замещения молекулы триазина.[10] Результаты эксперимента показывают, что возможно удалить 46 из 60 -ти исходных независимых переменных регрессии без потери качества предсказания.

В заключение автор выражает глубокую благодарность своему научному руководителю А. С. Антипину, учителю и научному консультанту диссертации И.Б. Мучнику, другу и коллеге по научной работе Л.В. Шварцеру за постоянное внимание и помощь в работе.

Список публикаций по теме диссертации

1. Гончаров Ю.В., Мучник И.Б., Шварцер Л.В. Алгоритм выбора признаков в задаче обучения классификации методом опорных векторов – Докл. 13-й всероссийской конф. Математические методы распознавания образов (ММРО-13). –М: ООО «МАКС Пресс». – 2007. – 700 с.
2. Гончаров Ю.В., Мучник И.Б., Шварцер Л.В. Алгоритм выбора признаков в задаче обучения классификации методом опорных векторов // Ж. вычисл. матем. и матем. физ. –2008. – Т.48. №.7. – С.1318-1336.
3. Гончаров Ю.В. Минимаксная задача выбора признаков для построения классификатора методом опорных векторов // Ж. вычисл. матем. и матем. физ. –2010. – Т.50. №.5. – С.967-976.
4. Goncharov Y., Muchnik I., Shvartser L. Simultaneous feature selection and margin maximization using saddle point approach // DIMACS Technical Report 2004. – №2004-08. – P.54.
5. Goncharov Y., Muchnik I., Shvartser L. Saddle point feature selection in SVM regression // DIMACS Technical Report 2007. – №2007-08. – P.15.
6. Goncharov Y., Muchnik I., Shvartser L. Saddle point feature selection in SVM classification // DIMACS Technical Report 2007. – №2007-16. – P.23.

[10] King R., Hirst J., Sternberg M. Comparison of artificial intelligence methods for modeling pharmaceutical QSARS // Applied Artificial Intelligence: An International Journal. 1995. V.9. Issue 2. P.213–233.