На правах рукописи

Лысёнок Евгений Игоревич

Построение базиса на множестве алгоритмов, основанных на гиперплоскостях, для произвольной задачи распознавания

Специальность 01.01.09 — дискретная математика и математическая кибернетика

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва
2010

Работа выполнена в организации Московский Государственный Университет имени М.В. Ломоносова, факультет Вычислительной Математики и Кибернетики.

Научный руководитель:	доктор физико-математических наук, академик РАН Журавлёв Юрий Иванович.
Официальные оппоненты:	доктор физико-математических наук, Сметанин Юрий Геннадиевич, кандидат физико-математических наук Кольцов Петр Петрович.
Ведущая организация:	Московский Физико-Технический Институт, факультет управления и прикладной математики.

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

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

Автореферат разослан _____

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

Общая характеристика работы

Актуальность темы. Проблема автоматического распознавания образов является одной из актуальных проблем математической кибернетики. Эта проблема является ведущей в таких областях науки и техники, как практическая геология, биология, химия, медицина и т.п. Одной из причин распространения данной проблемы в этих областях является то, что для применения методов распознавания требуется значительно меньшая точность описываемых объектов и явлений, чем при применении других методов прикладной математики. Второй важной причиной является то, что идея прецедентности, то есть идея принятия обоснованного решения на основе сравнения с другими объектами или явлениями, является ключевой для естественных наук.

Для решения задач распознавания было предложено множество методов. В том числе алгоритмы нахождения минимального эмпирического риска (Вапник В.Н. Червоненкис А.Я), метод потенциальных функций (йзерман М.А., Броверман Э.М., Розенэр Л.И.), алгоритмы вычисления оценок (Журавлёв Ю.И., Никифоров В.В.). Ю.И. Журавлёвым введено понятие алгебраического подхода к задачам распознавания, позволяющего синтезировать результаты распознавания алгоритмов различного рода. Было доказано, что с помощью алгебраических операций над некоторыми семействами алгоритмов можно построить алгоритм, точно распознающий любую наперёд заданную выборку при заданной начальной информации. Из каждого семейства, обладающего этим свойством, можно выделить конечный базис алгоритмов, который достаточен для данного построения.

Цель работы состоит в разработке метода построения базиса для семейства алгоритмов, основанных на разделяющих гиперплоскостях, и исследовании свойств полученного алгоритма.

В настоящей работе доказано существование данного базиса для

любых обучающих векторов и информации о классах и описан алгоритм его построения. Построение базиса также показано на примерах и разработан метод оптимизации получаемых с помощью базиса алгоритмов.

Методика исследований. Для доказательства использован математический аппарат комбинаторики, линейной алгебры, математической логики. Для разработки оптимизационного метода использован экспериментально-теоретический подход. Проведены эксперименты на данных реальных прикладных задач.

Научная новизна. Все результаты, полученные в диссертации, являются новыми. В диссертации доказано утверждение, ранее рассматриваемое лишь для ограниченного класса задач и более широкого множества алгебраических операций. Описан алгоритм построения базиса и предложен подход для улучшения качества получаемых с помощью базиса алгоритмов на новых выборках.

Апробация работы. Результаты, изложенные в диссертации, представлены на научных семинарах ВЦ РАН и кафедры Математических Методов Прогнозирования МГУ.

Основные результаты диссертации.

1) Доказана теорема о корректности линейного замыкания семейства алгоритмов, основанных на гиперплоскостях.

2) Создана программа для построения базиса исследуемого семейства.

3) Предложен подход для оптимизации построения базиса.

4) Реализован программный модуль, позволяющий применять указанный подход на данных различных форматов.

5) Исследуемые методы применены на данных прикладных задач.

Практическая и теоретическая ценность. Приведены теоретические результаты, касающиеся семейств распознающих алгоритмов, основанных на разделении гиперплоскостями. Показано, что с помощью линейных комбинаций можно построить корректное семейство. Таким образом, показано, что для достижения корректности данного семейства, нет необходимости расширять базовые алгоритмы или вводить новые операции. Результат и доказательство даёт возможность получать аналогичные утверждения для других семейств алгоритмов, что может влиять на их дальнейшее развитие и разработку.

Подход, используемый при доказательстве основной теоремы диссертации, позволяет явно выписывать корректный базис семейства. Это продемонстрировано на примерах реальных задач.

Предложен метод для улучшения качества получаемых с помощью базиса алгоритмов, что позволяет эффективно решать ряд прикладных задач из других областей науки.

Публикации. По теме диссертации опубликованы две работы в ЖВМиМФ (без соавторов).

Структура и объём работы. Работа состоит из введения, трёх глав и списка литературы из 26 наименований. Общий объём работы — 93 страницы, включая 22 рисунка.

Содержание диссертационной работы

Во **введении** даётся обзор различных классов алгоритмов распознавания. Приводятся ранее полученные результаты для семейств алгоритмов, родственных исследуемым.

В **первой главе** вводятся основные определения и обозначения, рассматривается постановка задачи, даётся понятие алгоритма, основанного

на разделяющей гиперплоскости.

Оператор R_A называется *распознающим оператором*, если он переводит

$$(I_0(K_1,\ldots,K_l), I(S_1,\ldots,S_q)), I_0(K_1,\ldots,K_l) \in I$$

в числовую матрицу $\{a_{ij}\}_{q\times l} = M_{q\times l}$.

Оператор r_A называется *решающим правилом*, если он переводит произвольную числовую матрицу $\{a_{ij}\} = M_{q\times l}$ в информационную матрицу

$$\{\alpha_{ij}^A\}_{q\times l} = I_{q\times l},$$

то есть матрицу, составленную из элементов $\{0, 1, \Delta\}$.

Решающее правило $\widetilde{r_A}$ называется *корректным*, если для всякой конечной совокупности допустимых объектов $S_1', \ldots, S_q'$ существует числовая матрица $\{\alpha_{ij}\}_{q\times l}$ такая, что $\widetilde{r_A}$ переводит $\{a_{ij}\}$ в матрицу, истинную для $S_1', \ldots, S_q'$.

Оператор $\widetilde{A}$ являющимся произведением распознающего оператора R_A и решающего правила r_A называется *стандартным распознающим алгоритмом.*

Пусть заданы элемент $I_0(l)$ из $\{I_0\}$,описания $I_S(q)$ допустимых объектов $S_1', \ldots, S_q'$, а также предикаты $P_1, \ldots, P_l$ на множестве допустимых объектов $\{S\}$, определяющих вхождение векторов $S_1, \ldots, S_q$ в классы $K_1, \ldots, K_l$. *Задача* $Z = Z(I_0, S_1, \ldots, S_q, P_1, \ldots, P_l)$ состоит в построении алгоритма для вычисления по I_0 свойств $P_1, \ldots, P_l$ для объектов $S_1, \ldots, S_q$.

Распознающий алгоритм A называется *корректным для задачи* Z, если для задачи Z выполнено условие $A(I_0(l), I(S_1, \ldots, S_q)) = \{\alpha_{ij}\}_{q\times l}$, где $\{\alpha_{ij}\}_{q\times l}$ — истинная информационная матрица для объектов $S_1, \ldots, S_q$. Семейство алгоритмов $\{A\}$ называется *корректным над множеством задач* $\{Z\}$, если для любой задачи $Z \in \{Z\}$ найдётся алгоритм $A \in \{A\}$, являющийся корректным для Z.

Пусть $B', B'' \in \{R_A\}$ — два произвольных распознающих оператора.

$B'(I_0, S_1, \ldots, S_q) = \{a'_{ij}\}_{q\times l}, B''(I_0, S_1, \ldots, S_q) = \{a''_{ij}\}_{q\times l}, b'$-произвольное скалярное вещественное число. Операторы $b' \cdot B, B' + B'', B' \cdot B''$ определяются следующим образом:

$$b' \cdot B'(I, S_1, \ldots, S_q) = \{b' \cdot a'_{ij}\}_{q\times l}; \tag{1}$$

$$(B' + B'')(I, S_1, \ldots, S_q) = \{a'_{ij} + a''_{ij}\}_{q\times l}; \tag{2}$$

$$(B' \cdot B'')(I, S_1, \ldots, S_q) = \{a'_{ij} \cdot a''_{ij}\}_{q\times l}. \tag{3}$$

Замыкание $L\{B\}$ множества $\{B\}$ относительно операций (1)-(3) называется *алгебраическим замыканием распознающих операторов*. Соответствующее семейство распознающих алгоритмов с фиксированным корректным решающим правилом называется *алгебраическим замыканием распознающих алгоритмов*.

Замыкание $L\{B\}$ множества $\{B\}$ относительно операций (1)-(2) называется *линейным замыканием распознающих операторов*. Соответствующее семейство распознающих алгоритмов с фиксированным корректным решающим правилом называется *линейным замыканием распознающих алгоритмов*.

Рассматриваются описания объектов $\{S\} \in M = M_1 \times \cdots \times M_n, M_i \subset \mathbb{R}$ в виде вещественных векторов $(a_1, a_2, \ldots, a_n), a_i \in M_i$ размерности n, то есть *стандартная обучающая информация*.

Пусть $\{R\}$ — совокупность кусочно-линейных поверхностей в n-мерном евклидовом пространстве.

Алгоритм A, основанный на кусочно-линейной поверхности, является алгоритмом распознавания для стандартной обучающей информации и определяется заданием кусочно-линейной поверхности R, набора неотрицательных числовых параметров $\gamma(S'_i), i = 1, \ldots, m$, параметров $x^{\delta}_{\alpha\beta}$,

задающих вхождение слагаемых в выражении для результата распознающего оператора, и корректного решающего правила, переводящего числовую матрицу $q \times l$ в булеву матрицу той же размерности.

Пусть R — гиперповерхность в n-мерное числовом пространстве, разбивающая его на два подмножества. Для объектов S, входящих в одно из подмножеств и таких, что $R(S) \neq 0$, будем писать $R(S) > 0$. Для объектов S, входящих в другое подмножество, если $R(S) \neq 0$, то будем писать $R(S) < 0$.

Алгоритм A представляет собой произведение распознающего оператора B и корректного решающего правила C^*. B зависит от параметров $(R, \widetilde{\gamma}^m, \widetilde{x})$. Оператор B по выборке $S_1, \ldots, S_q$ строит числовую матрицу:

$$\|a_i(S_j)\|_{q\times l} = \|a_{ij}\|_{q\times l}.$$

Для произвольного допустимого объекта строка $B(S) = (a_1(S), \ldots, a_l(S))$ определяется следующим образом.

1. $R(S) = 0$. Тогда $a_i(S) \equiv const$, $a_i(S)$ не зависит от параметров алгоритма: разделяющей поверхности R и параметров $\gamma_1, \ldots, \gamma_m, x^{\delta}_{\alpha\beta}$.

2. $R(S) \neq 0$. Разобьём объекты $S'_1, \ldots, S'_m$ такие, что $R(S'_i) \neq 0$ на подмножества $\mathfrak{M}^j_{\alpha,\beta}, \alpha = 0, 1, \beta = 0, 1, j = 1, \ldots, l$. Здесь:

α — значение предиката $P_j(S'_i) :< S'_i \in K_j >, 1 \leqslant j \leqslant l$;

β — значение предиката $Q(S'_i) :< R(S'_i) > 0 >$.

Множество $\mathfrak{M}^j_{\alpha\beta}$ состоит из таких S'_i, что $R(S'_i) \neq 0, P_j(S'_i) = \alpha, Q(S'_i) = \beta$.

Положим $\Gamma^j_{\alpha,\beta} = \sum_{S'_i \in \mathfrak{M}^j_{\alpha\beta}} \gamma(S'_i)$. Если $\{\mathfrak{M}\}^j_{\alpha,\beta}$ пусто, то $\Gamma^j_{\alpha\beta} = 0$.

1°. $R(S) > 0, B(S) = (a_1(S), \ldots, a_l(S))$,

$$a_j(S) = \frac{x^0_{11} \cdot \Gamma^j_{11} + x^0_{00} \cdot \Gamma^j_{00}}{x^0_{01} \cdot \Gamma^j_{01} + x^0_{10} \cdot \Gamma^j_{10} + 1}. \tag{4}$$

2°. $R(S) < 0$,

$$a_j(S) = \frac{x^1_{01} \cdot \Gamma^j_{01} + x^1_{10} \cdot \Gamma^j_{10}}{x^1_{11} \cdot \Gamma^j_{11} + x^1_{00} \cdot \Gamma^j_{00} + 1}. \tag{5}$$

В

(4) и (5) область изменения параметров γ_i есть $(0, +\infty), i = 1, \ldots, m$, параметры $x^{\delta}_{\alpha,\beta}$ принимают значение $0, 1$.

Рассмотрим класс $\widetilde{Z}$ задач $\widetilde{Z} =< I, S_1, \ldots, S_q >=< S'_1, \widetilde{\alpha}(S'_1); \ldots; S'_m, \widetilde{\alpha}(S'_m); S_1, \ldots, S_q >$, удовлетворяющих следующим условиям:

1) $\{S'_1, \ldots, S'_m\} \cap \{S_1, \ldots, S_q\} = \varnothing$

2) матрица $\|a_{ij}\|_{m \times l}$, строками которой являются информационные векторы объектов $S'_1, \ldots, S'_m$, не содержит одинаковых столбцов.

Совокупность задач $\widetilde{Z}$, удовлетворяющих условиям 1) и 2) обозначим через $\{\widetilde{Z}\}$.

Условия 1) и 2) являются естественными. Условие 1) означает, что $S_i \in \{S'_1, \ldots, S'_m\}$ и задача распознавания решена, так как известна принадлежность классам $K_1, \ldots, K_l$. Нарушение условия 2) означает, что два класса на обучающей выборке неразличимы.

Положим $\widetilde{M} = \{S'_1, \ldots, S'_m\}$.

Класс K_j называется изолированным в $\widetilde{Z}$, если существуют классы $K_r, K_t, K_v, 1 \leqslant j, r, t, v \leqslant l$ такие, что:

1) $K_j \bigcap \widetilde{M} \subset K_r \bigcap \widetilde{M}$;

2) $K_t \bigcap \widetilde{M} \subset K_j \bigcap \widetilde{M}$;

3) $(K_j \bigcap K_v) \bigcap \widetilde{M} = \varnothing$ и $\widetilde{M} \bigcap (K_j \bigcup K_v) = \widetilde{M}$.

Рассмотрим подкласс задач $\{Z\} \subset \{S\}$, состоящих из задач, не содержащих изолированных классов.

Следующие две теоремы были доказаны в более ранних работах.

Теорема 1.1. *Линейное замыкание $L\{A\}$ класса алгоритмов, основанных на кусочно-линейных поверхностях, является корректным на $\{Z\}$.*

В алгебре $\mathfrak{A}\{B\}$ распознающих операторов, соответствующих алгоритмам A, основанным на кусочно-линейных поверхностях, выделим операторы, представимые полиномами степени не выше 2. Множество таких операторов обозначим $\mathfrak{A}_2\{B\}$, а множество алгоритмов

$A = B' \cdot C, B' \in \mathfrak{A}_2\{B\}$, (C^* — фиксированное корректное решающее правило) — через $\mathfrak{A}_2\{A\}$.

Теорема 1.2. *Совокупность алгоритмов $\mathfrak{A}_2\{A\}$. корректна на множестве задач $\{\widetilde{Z}\}$.*

Во **второй главе** диссертации доказывается утверждение, являющееся усилением приведённых теорем:

Теорема 2.3. *(О корректности алгоритмов на основе гиперплоскостей). Линейное замыкание $L\{A\}$ класса алгоритмов $\{A\} = \{B \cdot C^*\}$ с произвольным корректным решающим правилом C^* и операторами $B \in \{B(\widetilde{x}, \widetilde{\gamma}, H)\}$, где $H \in \{R\}$ является гиперплоскостью, является корректным на $\widetilde{Z}$.*

Доказательство теоремы опирается на три леммы с номерами 1,3 и 4. При этом лемма 3 опирается на промежуточное утверждение, обозначенное леммой 2.

Лемма 1 (О разделяющем семействе гиперплоскостей).

Пусть дано произвольное множество из k различных точек в вещественном пространстве $\{S\} = \{S_1, \ldots, S_k\} \in \mathbb{R}^n$. Тогда существует система из k гиперплоскостей:

$$H_1, \ldots, H_k, \tag{6}$$

таких, что

$$1.|H_1^- \cap \{S\}| = 1, \ldots, |H_k^- \cap \{S\}| = k,$$

$$2.H_1^- \subset H_2^- \subset \cdots \subset H_k^-$$

$$3.H_1^+ \supset H_2^+ \supset \cdots \supset H_k^+,$$

$$4.H_i \cap \{S\} = \varnothing, i = 1, 2, \ldots, k.$$

Лемма 2 (О замкнутости линейных подпространств в конечномерном вещественном линейном пространстве). Любая предельная точка

подпространства H конечномерного вещественного пространства L принадлежит подпространству H.

Лемма 3 (О полноте линейного замыкания частного линейных комбинаций).

Пусть $\{a_{ij}\}_{n\times m}$ — булева матрица размеров $n \times m, a_{ij} \in \{0,1\}, i = 1,2,\ldots,m, j = 1,2,\ldots,n$ b — булев вектор размера $m, b_i \in \{0,1\}, i = 1,2,\ldots,m$. Обозначим через $a^1,\ldots,a^n$ столбцы матрицы $\{a_{ij}\}$:

$$a^j = \begin{pmatrix} a_{1j} \\ \ldots \\ a_{mj} \end{pmatrix}.$$

Пусть дано отображение M из декартова произведения множеств вещественных векторов размера m с положительными координатами, булевых векторов размера m и булевых матриц размеров $n \times m$ во множество вещественных матриц с двумя строками и n столбцами, определяемое соотношениями

$$M : \mathbb{R}^m_+ \times B^m \times B^{m\times n} \longrightarrow \mathbb{R}^{2\times n},$$

$$M(\gamma, b, A) = \begin{pmatrix} \mu_{11} & \ldots & \mu_{1n} \\ \mu_{21} & \ldots & \mu_{2n} \end{pmatrix}, i = 1,\ldots,6, \tag{7}$$

$$\mu_{1j} = \frac{< \gamma, a^j >}{1+ < \gamma, \overline{a^j} >},$$

$$\mu_{2j} = \frac{< \gamma \otimes b, \overline{a}^j >}{1+ < \gamma, a^j >}$$

для произвольных $x \in \mathbb{R}^m_+, b \in B^m, A \in B^{m\times n}$.

Для фиксированных матрицы A и столбца b пусть даны отображения M_1, M_2, M_3 из множества вещественных векторов длины m с положительными координатами во множество вещественных матриц с двумя

строками и n столбцами

$$M_i : \mathbb{R}_+^m \longrightarrow \mathbb{R}^{2\times n},\ i = 1, 2, 3, \tag{8}$$

определяемые следующими соотношениями:

$$M_1(\gamma) = M(\gamma, \sigma^m, A),\ M_2(\gamma) = M(\gamma, b, A),\ M_3(\gamma) = M(\gamma, \overline{b}, A), \tag{9}$$

где M определяется соотношениями (7).

Пусть фиксированы матрица A из $B^{m\times n}$, у которой все столбцы различны, и произвольный булев вектор b из B^m.

Тогда линейное замыкание объединения образов отображений (9) совпадает с подпространством матриц из двух строк длины n вида

$$L' = \{\begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \end{pmatrix} | \\ x_{1j}, x_{2j} \in \mathbb{R}, 1 \leqslant j \leqslant n, x_{1c_0} = 0, x_{2c_1} = 0\}, \tag{10}$$

где соотношение $x_{1c_0} = 0$ должно быть выполнено, если существует c_0 — номер нулевого столбца в матрице A, соотношение $x_{2c_1} = 0$ должно быть выполнено, если существует c_1 — номер столбца матрицы A, содержащего только единицы.

Лемма 4 (О конструировании базиса матриц размерности $m \times n$ на основе базиса двустрочных матриц). Пусть задан базис

$$E^{11}, \dots, E^{1n}, E^{21}, \dots, E^{2n} \tag{11}$$

пространства матриц $\mathbb{R}^{2\times n}$. Пусть элементы матрицы $E^{ij}, i = 1, 2, j = 1, 2, \dots n$ равны E^{ij}_{st}.

Пусть задана система из m булевых векторов $b^1, \dots, b^m$ такая что $b^p \Rightarrow b^r = 1, b^p \neq b^r, 1 \leqslant p < r \leqslant m$, т.е. координаты с единицами вектора b^p содержатся во множестве координат единиц следующего за

ним вектора b_{p+1}, не равного $b^p, 1 \leqslant p < n$. Очевидно, что при таком условии имеем $b^n = \sigma^n$, где σ^n, где вектор из n единиц.

Для каждого фиксированного вектора $b^p, p = 1, 2, \ldots, m$ с помощью системы (11) построим систему матриц из $\mathbb{R}^{m\times n}$:

$$E^{11,p} = \begin{pmatrix} E^{11}_{\mu_1 1} & \ldots & E^{11}_{\mu_1 n} \\ \ldots & & \\ E^{11}_{\mu_i 1} & \ldots & E^{11}_{\mu_i n} \\ \ldots & & \\ E^{11}_{\mu_m 1} & \ldots & E^{11}_{\mu_m n} \end{pmatrix},$$

$$\ldots\ldots\ldots$$

$$E^{2n,p} = \begin{pmatrix} E^{2n}_{\mu_1 1} & \ldots & E^{2n}_{\mu_1 n} \\ \ldots & & \\ E^{2n}_{\mu_i 1} & \ldots & E^{2n}_{\mu_i n} \\ \ldots & & \\ E^{2n}_{\mu_m 1} & \ldots & E^{2n}_{\mu_m n} \end{pmatrix}, \tag{12}$$

где

$$\mu_i = \begin{cases} 1, & b^p_i = 1, \\ 2, & b^p_i = 0, \end{cases}$$

т.е. строка с номером i у матрицы $E^{ij,p}$ равна первой строке матрицы базиса двустрочных матриц E^{ij} в случае, если i-й элемент вектора b^p равен 1, и второй строке той же матрицы в обратном случае, т.е. в случае 0 в i-й координате b^p. Тогда система матриц (12) полна в $\mathbb{R}^{m\times n}$.

В **третьей главе** приводятся результаты экспериментальных исследований и предлагается подход для улучшения качества рассматриваемых алгоритмов.

Для демонстрации конструктивности доказательства основного утверждения статьи приведено описание программы, которая для произвольной начальной информации и произвольных объектов распознавания конструирует базис на множестве алгоритмов, основанных на гиперповерхностях, с помощью которого можно получить произвольную заданную матрицу оценок.

Приведены примеры, как от выбора начального вектора при построении семейства гиперплоскостей, может зависеть качество распознавания на новых проверочных векторах. Учитывая этот факт, для улучшения качества распознавания алгоритма, полученного с помощью теоремы о корректности предлагается выделить из исходного множества объектов с известным результатом принадлежности к классам, множество векторов для проверки, и провести несколько итераций построения алгоритма с выбором разных начальных векторов для коэффициентов гиперплоскостей.

Для реализации предложенного подхода была создана программа, с помощью которой была показана эффективность подхода на практике. В качестве входных данных использованы данные предсказания двухлетней выживаемости больных остиогенной саркомой.

В **заключении** приводится краткий итог диссертации.

Таким образом, на зашиту выносится:

1) Теорема о корректности линейного замыкания алгоритмов на основе гиперплоскостей.

2) Метод построения базиса алгоритмов на основе гиперплоскостей.

3) Метод улучшения качества полученных алгоритмов за счёт оптимизации выбора начального вектора.

Список публикаций по теме диссертации

1. Лысёнок Е.И. Корректность линейного замыкания распознающих алгоритмов, основанных на гиперплоскостях. ЖВМиМФ. т. 49, № 10, С. 1885–1904, 2009.
2. Лысёнок Е.И. О некотором подходе выбора гиперплоскости для алгоритмов распознавания. ЖВМиМФ. т. 50, № 10, С. 1862–1864, 2010.