На правах рукописи

БОТОВ ПАВЕЛ ВАЛЕНТИНОВИЧ

ОЦЕНКИ ВЕРОЯТНОСТИ ПЕРЕОБУЧЕНИЯ МНОГОМЕРНЫХ СЕМЕЙСТВ АЛГОРИТМОВ КЛАССИФИКАЦИИ

05.13.17 — теоретические основы информатики

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2011

Работа выполнена на кафедре интеллектуальных систем Московского физико-технического института (государственного университета)

Научный руководитель:	доктор физико-математических наук, **Воронцов Константин Вячеславович**
Официальные оппоненты:	доктор физико-математических наук, академик РАН, **Матросов Виктор Леонидович**
	кандидат физико-математических наук, **Докукин Александр Александрович**
Ведущая организация:	Московский государственный университет им. М. В. Ломоносова, факультет ВМК

Защита диссертации состоится «22» декабря 2011 г. в _____ на заседании диссертационного совета Д 002.017.02 в Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

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

Автореферат разослан « ____ » ноября 2011 г.

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

Общая характеристика работы

Диссертационная работа относится к математической теории распознавания и классификации и посвящена проблеме повышения обобщающей способности алгоритмов классификации с помощью точных комбинаторных оценок вероятности переобучения для модельных семейств алгоритмов.

Актуальность темы. Проблема переобучения является одной из важнейших при решении задач восстановления зависимостей по эмпирическим данным. Она заключается в том, что частота ошибок на независимой контрольной выборке как правило, оказывается несколько выше частоты ошибок на обучающей выборке. Получением количественных оценок обобщающей способности, то есть вероятностного распределения величины этого смещения, занимается статистическая теория обучения. Многие подходы, разработанные в рамках этой теории, дают сильно завышенные оценки переобучения, поэтому их практическое применение не всегда приводит к улучшению качества восстановления зависимости. В комбинаторной теории надёжности обучения по прецедентам, предложенной К. В. Воронцовым, показано, что для получения точных оценок необходимо совместно учитывать свойства расслоения и связности семейства алгоритмов. Под расслоением семейства имеется в виду распределение алгоритмов по частоте ошибок, порождаемое заданной выборкой. Связность предполагает, что для каждого алгоритма в семействе найдётся множество похожих алгоритмов, отличающихся от него только на одном объекте выборки. Семейства, не обладающие свойствами расслоения и связности, могут переобучаться настолько сильно, что их практическое применение становится нецелесообразным.

Точные комбинаторные оценки вероятности переобучения были ранее получены для модельных семейств алгоритмов — монотонных и унимодальных цепей, интервалов, шаров и слоёв булева куба, обладающих тем или иным видом симметрии. Реальные семейства, порождаемые практическими задачами, как правило, имеют более сложную нерегулярную структуру, что препятствует получению точных комбинаторных оценок. Кроме того, согласно существующим в статистической теории обучения представлениям, размерность или число свободных параметров, как мера сложности семейства, определяющим образом влияет на его обобщающую способность. Поэтому актуальной теоретической проблемой является построение и изучение таких модельных семейств, которые обладали бы расслоением, связностью, размерностью, несимметричностью, то есть всеми ключевыми свойствами реальных семейств, и могли бы использоваться для их аппроксимации. Актуальной практической проблемой является применение оценок переобучения для повышения качества решения прикладных задач распознавания и классификации.

Цель работы: получение комбинаторных оценок обобщающей способности многомерных модельных семейств алгоритмов, разработка методов аппроксимации вероятности переобучения реальных семейств вероятностью переобучения модельных семейств, применение этих оценок для повышения обобщающей способности итерационных методов обучения в задачах классификации.

Научная новизна. Предложены и исследованы модельные семейства алгоритмов — связки монотонных цепей, монотонные и унимодальные сети, обладающие свойствами расслоения, связности, размерности и несимметричности. Для всех семейств полу-

чены точные комбинаторные формулы вероятности переобучения и математического ожидания частоты ошибок на генеральной выборке. Предложен метод минимизации предсказанного риска (МПР), основанный на замене реального семейства подходящим по структуре модельным семейством, с последующей минимизацией ожидаемой частоты ошибок на генеральной выборке. В отличие от метода минимизации структурного риска, МПР учитывает особенности конкретной выборки. В отличие от скользящего контроля, МПР не требует многократного обучения, и потому вычислительно гораздо более эффективен. Предложена общая методика применения МПР в итерационных методах обучения, показано её применение на примере решающих деревьев.

Методы исследования. Для получения оценок вероятности переобучения использовалась перестановочная вероятностная аксиоматика, комбинаторная теория надёжности обучения по прецедентам, элементы комбинаторики и теории вероятностей. Для проверки точности оценок проводились численные эксперименты на модельных данных методом Монте-Карло. Для сравнения предлагаемых методов классификации со стандартными проводились эксперименты на реальных данных из репозитория UCI.

Положения, выносимые на защиту.

1. Методы получения комбинаторных оценок вероятности переобучения на основе послойного разложения семейства.

2. Оценки вероятности переобучения модельных семейств алгоритмов: связки цепей, монотонной сети, унимодальной сети и их несимметричных аналогов.

3. Методика повышения обобщающей способности итерационных методов обучения с помощью комбинаторных оценок вероятности переобучения.

Теоретическая и практическая значимость. Данная работа вносит существенный вклад в развитие комбинаторной теории надёжности обучения по прецедентам.

Метод послойного разложения семейства (глава 2) может быть применён для получения оценок вероятности переобучения в более широком классе семейств.

Метод минимизации предсказанного риска (глава 4) может быть применён для повышения качества классификации в широком классе итерационных методов обучения, включая решающие деревья, решающие списки, композиции логических закономерностей, алгоритмы вычисления оценок, и другие.

Апробация работы. Результаты работы докладывались, обсуждались и получили одобрение специалистов на следующих научных конференциях и семинарах:

- Всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [1];
- Международная конференция «Интеллектуализация обработки информации» ИОИ-8, 2010 г. [2];
- Всероссийская конференция «Математические методы распознавания образов» ММРО-15, 2011 г. [4];
- Результаты работы неоднократно докладывались на семинарах отдела Интеллектуальных систем ВЦ РАН.

Публикации по теме диссертации. Всего публикаций по теме диссертации — четыре, в том числе в изданиях из списка, рекомендованного ВАК РФ — одна [3].

Структура и объём работы. Работа состоит из введения, четырёх глав, заключения, приложения, списка используемых источников, состоящего из 54 наименований. Общий объем работы составляет 99 страниц.

Краткое содержание работы по главам

Глава 1. Комбинаторная теория переобучения

1.1. Основные понятия и обозначения. Пусть задано множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, называемое *генеральной выборкой*, конечное *семейство алгоритмов* A и бинарная функция ошибки $I\colon A \times \mathbb{X} \to \{0, 1\}$. Если $I(a, x) = 1$, то говорят, что алгоритм a ошибается на объекте x. *Вектором ошибок* алгоритма a называется бинарный вектор $\big(I(a, x_i)\big)_{i=1}^{L}$. *Число ошибок* алгоритма a на выборке $X \subseteq \mathbb{X}$ есть $n(a, X) = \sum\limits_{x \in X} I(a, x)$. *Частота ошибок* алгоритма a на выборке $X \subseteq \mathbb{X}$ есть $\nu(a, X) = n(a, X)/|X|$.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной обучающей выборке $X \subseteq \mathbb{X}$ ставит в соответствие некоторый алгоритм. Частоту ошибок $\nu(\mu X, X)$ называют *эмпирическим риском*. Метод обучения μ называется минимизацией эмпирического риска (МЭР), если $\mu X \in A(X) = \operatorname{Arg}\min\limits_{a \in A} n(a, X)$ для любой обучающей выборки X. Метод МЭР называется *пессимистичным*, если $\mu X \in \operatorname{Arg}\max\limits_{a \in A(X)} n(a, \bar{X})$.

В задачах *обучения по прецедентам* требуется по выборке X построить алгоритм $a \in A$, который как можно реже ошибается не только на известных (наблюдаемых) объектах из X, но и на неизвестных (скрытых в момент обучения) объектах из $\bar{X}$.

Допустим, что все C_L^ℓ разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ на наблюдаемую обучающую выборку X длины ℓ и скрытую контрольную $\bar{X}$ длины $k = L - \ell$ равновероятны. *Вероятностью переобучения* метода μ на выборке $\mathbb{X}$ называется величина

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_X \big[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big].$$

Здесь и далее $\sum\limits_X$ означает суммирование по всем ℓ-элементным подмножествам $X \subset \mathbb{X}$ генеральной выборки; логическое выражение в квадратных скобках означает: $[\mathit{истина}] = 1$ и $[\mathit{ложь}] = 0$.

Если множество $A = \{a\}$ одноэлементно, то точное значение Q_ε задаётся функцией гипергеометрического распределения:

$$Q_\varepsilon = H_L^{\ell,m}(s_m(\varepsilon)), \quad \text{где} \quad H_L^{\ell,m}(z) = \sum_{s=s_0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell},$$

$s_m(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $s_0 = \max\{0, m - k\}$, $\lfloor z \rfloor$ — округление z до ближайшего целого, не большего z.

Глава 2. Новые методы получения комбинаторных оценок вероятности переобучения

2.1. Послойный метод. Множество $A_t = \{a \in A \colon n(a, \mathbb{X}) = t + m\}$ называется t-*слоем* множества алгоритмов A относительно выборки $\mathbb{X}$, где $m = \min\limits_{a \in A} n(a, \mathbb{X})$ — число ошибок *лучшего алгоритма*.

Теорема 1.1. *Если A — семейство алгоритмов, μ — метод обучения, векторы ошибок всех алгоритмов $a \in A$ попарно различны, то для каждого t-слоя A_t можно указать такое множество индексов V_t, и для каждого $v \in V_t$ такую пару непересекающихся подмножеств $X_{tv}, X'_{tv} \subset \mathbb{X}$ и такой коэффициент $c_{tv} \in \mathbb{R}$, что*

$$[\mu X \in A_t] = \sum_{v \in V_t} c_{tv} [X_{tv} \subseteq X][X'_{tv} \subseteq \bar{X}].$$

Гипотеза 1.1. Множества X_{tv} и X'_{tv} можно выбрать так, что $n(a, X_{tv}) = 0$ и $n(a, X'_{tv}) = t$ для всех $a \in A_t$.

Введём обозначения: $\ell_{tv} = \ell - |X_{tv}|,\ \ L_{tv} = L - |X_{tv}| - |X'_{tv}|$.

Теорема 1.2. *Пусть верны условия теоремы 1.1, тогда вероятность получить алгоритм из t-слоя:*

$$P_t = P[\mu X \in A_t] = \sum_{v \in V_t} c_{tv} \frac{C^{\ell_{tv}}_{L_{tv}}}{C^{\ell}_{L}}.$$

Теорема 1.3. *Пусть верны условия теоремы 1.1, а также верна гипотеза 1.1, тогда вероятность переобучения есть*

$$Q_\varepsilon = \sum_t Q_{\varepsilon,t},$$

где суммирование производится по всем слоям алгоритмов, а вклады t-слоёв в вероятность переобучения определяются как

$$Q_{\varepsilon,t} = \sum_{v \in V_t} c_{t,v} \frac{C^{\ell_{tv}}_{L_{tv}}}{C^{\ell}_{L}} H^{\ell_{tv},m}_{L_{tv}}(s_t(\varepsilon)).$$

Обозначим через n^* математическое ожидание числа ошибок на генеральной выборке: $n^* = m + \sum\limits_{t=0}^{k} tP_t$.

2.2. Метод индикаторов. Удобно ввести следующие определения для часто используемых бинарных выражений:

$[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}]$ — *элементарный индикатор*, где X_α и X_δ — пара из порождающего и запрещающего множеств;

$[\mu X = a]$ — *индикатор алгоритма* a;

$[\mu X \in A_t] = \sum\limits_{a \in A_t} [\mu X = a]$ — *индикатор* t*-слоя.*

Бинарное выражение $[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}]$ будем называть просто *индикатором*, если про X_α и X_δ не известно, являются ли они парой из запрещающего и порождающего множеств.

Согласно теореме 1.1, индикатор слоя может быть представлен как сумма элементарных индикаторов. Определим вклады элементарного индикатора в вероятность получения алгоритма a (или слоя A_t) и в вероятность переобучения семейства:

$$\mathsf{P}[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}] = \frac{C^{\ell-|X_\alpha|}_{L-|X_\alpha|-|X_\delta|}}{C^\ell_L},$$

$$Q_{\varepsilon,t}[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}] = \frac{C^{\ell-|X_\alpha|}_{L-|X_\alpha|-|X_\delta|}}{C^\ell_L} H^{\ell-|X_\alpha|,m}_{L-|X_\alpha|-|X_\delta|}\left(s_t(\varepsilon)\right).$$

Введём *редукцию* $\beta(a,b)$ как сокращённую запись индикатора $[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}]$, где $a = |X_\alpha|$, $b = |X_\delta|$, при условии, что $X_\alpha \cap X_\delta = \varnothing$.

Поскольку вклады зависят не от самих множеств X_α, X_δ, а только от их мощностей (a,b), удобно ввести обозначения:

$$P\beta(a,b) = P[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}] = \frac{C^{\ell-a}_{L-a-b}}{C^\ell_L};$$

$$Q_{\varepsilon,t}\beta(a,b) = Q_{\varepsilon,t}[X_\alpha \subseteq X][X_\delta \subseteq \bar{X}] = \frac{C^{\ell-a}_{L-a-b}}{C^\ell_L} H^{\ell-a,m}_{L-a-b}\left(s_t(\varepsilon)\right).$$

Две редукции $\beta(a,b)$ и $\beta(c,d)$ будем считать равными, если $P\beta(a,b) = P\beta(c,d)$ и $Q_{\varepsilon,t}\beta(a,b) = Q_{\varepsilon,t}\beta(c,d)$. Индикаторы этих редукций будем называть эквивалентными.

Некоторые свойства редукций.

1. Свойство суммы: $p\beta(a,b) + q\beta(a,b) = (p+q)\beta(a,b)$.
2. Свойство произведения: $\beta(a,b) \cdot \beta(c,d) = \beta(a+c, b+d)$.
3. Упрощение суммы: $\sum\limits_{i=a}^{b} \beta(c,i) = \beta(c-1,a) - \beta(c-1,b+1)$.

Основным приёмом доказательства оценок вероятности переобучения в следующей главе является представление индикатора слоя в виде взвешенной суммы элементарных индикаторов, которая упрощается с помощью операций над редукциями.

Глава 3. Модельные семейства алгоритмов

В данной главе вводятся модельные семейства алгоритмов, для них доказываются теоремы о вероятности переобучения Q_ε и о вероятности P_t выбора алгоритма из t-слоя, предлагаются эффективные алгоритмы вычисления величин Q_ε и P_t. Приводятся результаты численных экспериментов по методу Монте-Карло, показывающие, что величины Q_ε и P_t, вычисленные по полученным формулам, совпадают с точными значениями.

В данной главе модельные семейств вводятся по единой схеме.

1. Определение модельного семейства, пример бинарной матрицы «алгоритм» × «объект», диаграмма семейства;
2. Лемма о редукции индикатора слоя семейства алгоритмов;

3. Теоремы о вероятностях P_t выбора алгоритма из t-слоя и о вкладах $Q_{\varepsilon,t}$ слоёв в вероятность переобучения Q_ε;

Вероятности P_t необходимы для вычисления математического ожидания ошибки n^*, используемого в методе минимизации предсказанного риска в четвёртой главе.

3.1. Пара алгоритмов

Двухэлементное семейство изучено в работах К. В. Воронцова. В данной работе та же оценка получена методом индикаторов.

Теорема 1.4 (О паре алгоритмов). *Пусть $A = \{a_1, a_2\}$, m_1, m_2 и m_0 — число объектов, на которых, ошибается, соответственно, a_1, a_2 и оба алгоритма вместе. Пусть μ — пессимистическая минимизация эмпирического риска. Тогда вероятность при обучении выбрать алгоритмы a_1 и a_2 есть*

$$P_{a_1} = \sum_{i_2=0}^{m_2} \sum_{i_1=0}^{i_2} P(i_1, i_2), \qquad P_{a_2} = \sum_{i_2=0}^{m_2} \sum_{i_1=i_2+1}^{m_1} P(i_1, i_2),$$

где $P(i_1, i_2) = C_{m_1}^{i_1} C_{m_2}^{i_2} C_{L-m_1-m_2}^{\ell-i_1+i_2} / C_L^\ell$. Вероятность переобучения семейства есть $Q_\varepsilon = Q_{\varepsilon,a_1} + Q_{\varepsilon,a_2}$, где вклады каждого из алгоритмов

$$Q_{\varepsilon,a_1} = \sum_{i_2=0}^{m_2} \sum_{i_1=0}^{i_2} P(i_1, i_2) H_{L-m_1-m_2}^{\ell-i_1+i_2,m_0}\big(\tfrac{\ell}{L}(m_0 + m_1 - k\varepsilon) - i_1\big),$$

$$Q_{\varepsilon,a_2} = \sum_{i_2=0}^{m_2} \sum_{i_1=i_2+1}^{m_1} P(i_1, i_2) H_{L-m_1-m_2}^{\ell-i_1+i_2,m_0}\big(\tfrac{\ell}{L}(m_0 + m_2 - k\varepsilon) - i_2\big).$$

Заметим, что, несмотря на кажущуюся простоту семейства, формулы имеют весьма громоздкий вид.

3.2. Монотонная и унимодальная цепи

Точные оценки вероятности переобучения для монотонных и унимодальных цепей были получены ранее К. В. Воронцовым. В данной работе те же оценки выводятся послойным методом, что позволит в дальнейшем обобщить их на случай многомерных монотонных и унимодальных сетей.

3.2.1. Монотонная цепь. *Цепью алгоритмов* называется семейство алгоритмов, элементы которого можно выстроить в последовательность так, чтобы любые два соседних алгоритма различались только на одном объекте.

Цепь алгоритмов $A = \{a_0, a_1, a_2, \ldots, a_D\}$ называется *монотонной цепью* высоты D, если $n(a_d, \mathbb{X}) = m + d$ для всех $d = 0, \ldots, D$ при некотором $m \geqslant 0$. Алгоритм a_0 называется *лучшим в цепи*.

Цепь алгоритмов может порождаться однопараметрическим семейством классификаторов с непрерывной по параметру дискриминантной функцией. Монотонная цепь — это одно из простейших модельных семейств, обладающее свойствами расслоения и связности.

Теорема 1.5 (О вероятности выбора алгоритма). *Пусть A — монотонная цепь высоты D, μ — пессимистическая минимизация эмпирического риска и выполнено условие $D \leqslant k$. Тогда вероятность выбора алгоритмов из слоя t выражается в виде*

$$P_t = \frac{C_{L-t}^{\ell-\varkappa}}{C_L^\ell}, \quad \text{где} \quad \varkappa = [t = D].$$

Теорема 1.6 (О вероятности переобучения). *Пусть A — монотонная цепь высоты D, μ — пессимистическая минимизация*

эмпирического риска и выполнено условие $D \leqslant k$. *Тогда вероятность переобучения семейства есть* $Q_\varepsilon = \sum\limits_{t=0}^{D} Q_{\varepsilon,t}$, *вклады слоёв* $Q_{\varepsilon,t}$ *выражаются в виде*

$$Q_{\varepsilon,t} = \frac{C_{L-t}^{\ell-\varkappa}}{C_L^\ell} H_{L-t}^{\ell-\varkappa,m}(s_t(\varepsilon)),$$

где $s_t(\varepsilon) = \frac{l}{L}(m + t - \varepsilon k)$ *и* $\varkappa = [t = D]$.

3.2.2. Унимодальная цепь. Цепь $A = \{a_{-D}, \ldots, a_0, \ldots, a_{+D}\}$ называется *унимодальной цепью* алгоритмов высоты D, если

1. $n(a_d, \mathbb{X}) = m + |d|$ при некотором $m \geqslant 0$; алгоритм a_0 называется *лучшим в цепи* A;
2. если d и d' одного знака и $|d| < |d'|$, то $I(a_d, x) \leqslant I(a_{d'}, x)$ для всех $x \in \mathbb{X}$;
3. если d и d' имеют разные знаки, то существует лишь m объектов, на которых ошибаются оба алгоритма.

Унимодальная цепь является более адекватной моделью однопараметрического семейства алгоритмов, поскольку допускает отклонения от лучшего алгоритма в обе стороны.

Теорема 1.7 (О вероятности выбора алгоритма). *Пусть* A *— унимодальная цепь высоты* D, μ *— пессимистичная минимизация эмпирического риска и выполнено условие* $D \leqslant k$. *Тогда вероятность выбора алгоритма из слоя* t *выражается через*

$$P_t = \left(C_{L-2t-2\varkappa}^{\ell-2\varkappa} + 2C_{L-t-1\varkappa}^{\ell-1\varkappa} - 2C_{L-2t-1\varkappa}^{\ell-1\varkappa}\right)/C_L^\ell, \text{ где } \varkappa = [t < D].$$

Теорема 1.8 (О вероятности переобучения). *Пусть A — унимодальная цепь высоты D, μ — пессимистичная минимизация эмпирического риска, $\varkappa = [t < D]$, и выполнено условие $D \leqslant k$. Тогда вероятность переобучения семейства есть $Q_\varepsilon = \sum_{t=0}^{D} Q_{\varepsilon,t}$, где вклады $Q_{\varepsilon,t}$ выражаются через*

$$Q_{\varepsilon,t} = \Big(C_{L-2t-\varkappa}^{\ell-\varkappa} H_{L-2t-\varkappa}^{\ell-\varkappa,m}(s_t(\varepsilon)) + \\ + 2C_{L-t-\varkappa}^{\ell-\varkappa} H_{L-t-\varkappa}^{\ell-\varkappa,m}(s_t(\varepsilon)) - 2C_{L-2t-\varkappa}^{\ell-\varkappa} H_{L-2t-\varkappa}^{\ell-\varkappa,m}(s_t(\varepsilon)) \Big) / C_L^\ell.$$

3.3. Многомерные симметричные модели

Вектором индексов будем называть вектор $\mathbf{d} = (d_i)_{i=1}^h$ с целыми компонентами. Положим $|\mathbf{d}| = \sum_{i=1}^h |d_i|$. На множестве векторов индексов естественным образом введём частичный порядок: $\mathbf{d} < \mathbf{d}'$ тогда и только тогда, когда $d_i \leqslant d'_i$ для всех $i \in 1, \ldots, h$, и хотя бы одно из неравенств строгое. *Неотрицательным вектором индексов* будем называть вектор индексов $\mathbf{d}$ с $d_i \geqslant 0,\ i \in 1, \ldots, h$.

3.3.3. Монотонная сеть. Множество алгоритмов $A = \{a_\mathbf{d}\}$, где вектор индексов $\mathbf{d}$ с неотрицательными компонентами имеет размерность h и $|\mathbf{d}| \leqslant D$, называется *монотонной h-мерной сетью алгоритмов высоты D*, если выполнены условия:

1. если $\mathbf{d} < \mathbf{d}'$, то для всех $x \in \mathbb{X}$ выполнено $I(a_\mathbf{d}, x) \leqslant I(a_{\mathbf{d}'}, x)$, причём ровно $|\mathbf{d} - \mathbf{d}'|$ неравенств строгие;
2. $n(a_\mathbf{d}, \mathbb{X}) = m + |\mathbf{d}|$ при некотором $m \geqslant 0$ для всех $a_\mathbf{d} \in A$. Алгоритм $a_\mathbf{0}$ называется *лучшим*.

Монотонная сеть является моделью h-мерного параметрического семейства классификаторов с непрерывной по параметрам дис-

криминантной функцией. Это модельное семейство, обладающее одновременно свойствами расслоения, связности и размерности.

Теорема 1.9 (О вероятности выбора алгоритма). *Пусть A — h-мерная монотонная сеть высоты D, μ — пессимистичная минимизация эмпирического риска и выполнено условие $D \leqslant k$. Тогда вероятность выбора алгоритма из t-слоя,*

$$P_t = \begin{cases} C_{h+t-1}^{t} \frac{C_{L-h-t}^{\ell-h}}{C_L^{\ell}}, & t < D; \\ 1 - \sum\limits_{i=0}^{D-1} P_i, & t = D. \end{cases}$$

Теорема 1.10 (О вероятности переобучения). *Пусть A — h-мерная монотонная сеть высоты D, μ — пессимистичная минимизация эмпирического риска и выполнено условие $D \leqslant k$. Тогда вероятность переобучения семейства есть $Q_\varepsilon = \sum\limits_{t=0}^{D} Q_{\varepsilon,t}$ где*

$$Q_{\varepsilon,t} = \begin{cases} C_{h+t-1}^{t} \frac{C_{L-h-t}^{\ell-h}}{C_L^{\ell}} H_{L-h-t}^{\ell-h,m}(s_t(\varepsilon)), & t < D; \\ H_L^{\ell,m}(s_D(\varepsilon)) - \sum\limits_{i=0}^{D-1} C_{h+i-1}^{i} \frac{C_{L-h-i}^{\ell-h}}{C_L^{\ell}} H_{L-h-i}^{\ell-h,m}(s_D(\varepsilon)), & t = D. \end{cases}$$

В следующих параграфах аналогичные теоремы о вероятности выбрать алгоритм из t-слоя и о вероятности переобучения доказываются для других модельных семейств: связок монотонных цепей, унимодальных сетей, а также их несимметричных обобщений.

3.3.4. Связка цепей.

Определение 1.1. *Множество $\left\{a_d^b \colon b = 1, \ldots, h,\ d = 0, \ldots, D\right\}$ называется связкой h монотонных цепей высоты D, если*

1. *существует алгоритм $a^* \in A$, называемый лучшим в связке, такой, что $a^* = a_0^b$ для всех b;*

2. $n(a_d^b, \mathbb{X}) = m + d$ *при некотором* $m \geqslant 0$, *любом* $b \in \{1, \ldots, h\}$ *и любом* $d \leqslant D$;

3. *если* $d' > d$, *то* $I(a_d^b, x) \geqslant I(a_{d'}^b, x)$, *причём ровно* $d' - d$ *неравенств строгие;*

4. *если* $I(a_d^b, x) = I(a_{d'}^{b'}, x) = 1$, $b \neq b'$, *для некоторого объекта* x, *то* $I(a^*, x) = 1$.

Связка цепей является неполной моделью h-мерного параметрического семейства классификаторов с непрерывной по параметрам дискриминантной функцией. Она состоит только из алгоритмов, которые могут быть получены варьированием каждого из h параметров по отдельности в окрестности лучшего алгоритма.

3.3.5. Унимодальная сеть. Множество алгоритмов $A = \{a_{\mathbf{d}}\}$, где вектор индексов $\mathbf{d}$ имеет размерность h и $|\mathbf{d}| \leqslant D$, называется *унимодальной* h*-мерной сетью алгоритмов* высоты D, если выполнены условия:

1. если $\mathbf{0} \leqslant \mathbf{d} < \mathbf{d}'$ или $\mathbf{d}' < \mathbf{d} \leqslant \mathbf{0}$, то для всех $x \in \mathbb{X}$ выполнено $I(a_{\mathbf{d}}, x) \leqslant I(a_{\mathbf{d}'}, x)$, причём ровно $|d - d'|$ неравенств строгие.

2. $n(a_{\mathbf{d}}, \mathbb{X}) = m + |\mathbf{d}|$ при некотором $m \geqslant 0$ для всех $a_{\mathbf{d}} \in A$. Алгоритм $a_{\mathbf{0}}$ называется *лучшим*.

Унимодальная сеть отличается от монотонной снятием ограничения на неотрицательность компонент вектора индексов $\mathbf{d}$. Она является более адекватной моделью h-параметрического семейства классификаторов, поскольку допускает отклонения от лучшего алгоритма в обе стороны по каждому из параметров.

3.3.6. О некоторых свойствах многомерных семейств. Если для семейств алгоритмов A и A' при фиксированных L, ℓ, ε выполнены неравенства $n^*(A) \leqslant n^*(A')$ и $Q_\varepsilon(A) \leqslant Q_\varepsilon(A')$, то будем говорить, что семейство A' *мажорирует* A и записывать $A \prec A'$.

Теорема 1.11 (О мажорируемости). *Унимодальная сеть размерности h мажорирует связку цепей размерности $2h$ и мажорируется монотонной сеть размерности $2h$:*

(связка цепей)$_{2h}$ $\prec$ (унимодальная сеть)$_h$ $\prec$ (монотонная сеть)$_{2h}$.

Каждое из этих семейств мажорируется семействами большей размерности h или большей высоты D.

3.4. Несимметричные модели

Эксперименты показывают, что реальные семейства при варьировании одного из свободных параметров похожи на унимодальные цепи, но с ветвями разных высот. В данном разделе вводятся несимметричные монотонные сети, унимодальные сети и связки цепей, как естественное обобщение соответствующих симметричных модельных семейств.

3.4.1. Монотонная несимметричная сеть. Множество алгоритмов $A = \{a_{\mathbf{d}} \colon \mathbf{0} \leqslant \mathbf{d} \leqslant \mathbf{D},\ |\mathbf{d}| < D_0\}$, проиндексированное *вектором высот* $\mathbf{D}$ и общей высотой D_0, называется *монотонной несимметричной h-мерной сетью алгоритмов*, если

1. из $\mathbf{d} < \mathbf{d}'$ следует $I(a_{\mathbf{d}}, x) \leqslant I(a_{\mathbf{d}'}, x)$ для всех $x \in \mathbb{X}$, причём ровно $|d - d'|$ неравенств строгие;
2. $n(a_{\mathbf{d}}, \mathbb{X}) = m + |\mathbf{d}|$ при некотором $m \geqslant 0$ для всех $a_{\mathbf{d}} \in A$. Алгоритм $a_{\mathbf{0}}$ называется *лучшим*.

3.4.2. Несимметричная связка цепей. Множество алгоритмов $A = \{a_d^b \colon b = 1, \ldots, h,\ d = 0, \ldots, D_b\}$, называется несимметричной *связкой* h монотонных цепей высот $D_1, \ldots, D_h$, если

1. существует алгоритм $a^* \in A$, называемый *лучшим в связке*, такой, что $a^* = a_0^b$ для всех b;
2. $n(a_d^b, \mathbb{X}) = m + d$ при некотором $m \geqslant 0$, любом $b \in \{1, \ldots, h\}$ и любом $d \leqslant D_b$;
3. подмножество $A_b = \{a_d^b : d = 0, \ldots, D_b\}$ является монотонной цепью для всех $b \in \{1, \ldots, h\}$;
4. если $I(a_d^b, x) = I(a_{d'}^{b'}, x) = 1,\ b \neq b'$, для некоторого объекта x то $I(a^*, x) = 1$.

3.4.3. Унимодальная несимметричная сеть. Множество $A = \{a_{\mathbf{w}} \colon -\mathbf{W}^- \leqslant \mathbf{w} \leqslant \mathbf{W}^+,\ |\mathbf{w}| \leqslant W_0\}$, задаваемое двумя векторами высот $-\mathbf{W}^- \leqslant \mathbf{0} \leqslant \mathbf{W}^+$ и общей высотой W_0, называется *унимодальной несимметричной h-мерной сетью алгоритмов (УНС)*, если выполнены условия:

1. если $\mathbf{0} \leqslant \mathbf{w} < \mathbf{w}'$ или $\mathbf{w}' < \mathbf{w} \leqslant \mathbf{0}$, то для всех $x \in \mathbb{X}$ выполнено $I(a_{\mathbf{w}}, x) \leqslant I(a_{\mathbf{w}'}, x)$, причём ровно $|w - w'|$ неравенств строгие.
2. $n(a_{\mathbf{w}}, \mathbb{X}) = m + |\mathbf{w}|$ при некотором $m \geqslant 0$ для всех $a_{\mathbf{w}} \in A$. Алгоритм $a_{\mathbf{0}}$ называется *лучшим*.

3.4.4. О некоторых свойствах несимметричных многомерных семейств. Унимодальная несимметричная сеть является несимметричным обобщением ранее рассмотренной унимодальной

(симметричной) сети, которая, в свою очередь, является многомерным обобщением унимодальной цепи. С другой стороны, при $\mathbf{W}^- = \mathbf{0}$ УНС вырождается в несимметричную монотонную сеть.

УНС с векторами высот $\mathbf{W}^\pm$ мажорирует и мажорируется симметричными унимодальными сетями с высотами, соответственно $D_- = \min\limits_j \min(W_j^-, W_j^+)$ и $D_+ = \max\limits_j \max(W_j^-, W_j^+)$.

УНС с векторами высот $\mathbf{W}^\pm$ мажорируется УНС с векторами высот $\mathbf{W}^{\pm'}$, о которых известно, что $\mathbf{W}^{\pm'} > \mathbf{W}^\pm$.

Теорема 1.12 (О мажорируемости). *Несимметричная унимодальная сеть размерности h с векторами высот $\mathbf{W}^\pm = (W_j^\pm)_{j=1}^h$ мажорирует несимметричную связку цепей размерности $2h$ с высотами $W_1^-, W_1^+, \ldots, W_h^-, W_h^+$ и мажорируется несимметричной монотонной сетью размерности $2h$ с высотами $(W_1^-, W_1^+, \ldots, W_h^-, W_h^+)$ и общей для всех семейств главной высотой W_0:*

(связка цепей)$_{2h}$ $\prec$ (унимодальная сеть)$_h$ $\prec$ (монотонная сеть)$_{2h}$.

Глава 4. Вычислительные эксперименты на реальных данных

4.1. Аппроксимация семейств. Показано, что зависимости вероятности переобучения Q_ε от ε для четырёх стандартных семейств классификаторов (наивный байесовский, SVM, решающее дерево, нейронная сеть), порождаемые на реальных данных (5 реальных задачах из репозитория UCI), хорошо аппроксимируются зависимостями Q_ε от ε многомерных модельных семейств при подборе параметра размерности h. В частности, для задачи breast-cancer-wisconsin наилучшая аппроксимация получена для монотонной цепи, что согласуется с априорной информацией, что в данной задаче информативен только один признак.

Предложен эвристический метод подбора параметров m, h, $\mathbf{W}^{\pm}$ с целью получения унимодальной несимметричной сети, вероятности P_t и Q_ε которой могут быть использованы для аппроксимации вероятностей P_t и Q_ε реального семейства, порождаемого заданной выборкой (УНС-аппроксимация).

Эксперименты на реальных данных показали, что при УНС-аппроксимации оценка $n^* - m$ завышается на 5–10% для одномерных семейств, на 10–30% для двумерных и до 50% для трёхмерных. Более чем в половине случаев одномерные семейства аппрокимируются без ошибок, поскольку окрестность лучшего алгоритма радиуса 6 оказывается унимодальной цепью.

4.2. Метод минимизации предсказанного риска. Рассматривается широкий класс *итерационных методов обучения*, в которых процесс обучения состоит из последовательности шагов; на каждом шаге оптимизируется скалярный функционал, зависящий от выборки, например, эмпирический риск. Такие методы используются для обучения решающих деревьев, решающих списков, композиций логических закономерностей, алгоритмов вычисления оценок. Для улучшения обобщающей способности таких методов обучения, предлагается *метод предсказанного риска* (МПР), в котором на каждом шаге обучения минимизируется математическое ожидание ошибки на генеральной выборке n^*. МПР основан на предположении, что вероятность переобучения семейства алгоритмов, порождаемого обучающей выборкой X, не сильно отличается от вероятности переобучения семейства алгоритмов, порождаемого генеральной выборкой $\mathbb{X}$.

В экспериментах функционал n^* вычислялся двумя способами: эмпирически методом Монте-Карло и с помощью УНС-аппроксимации. Эксперименты на реальных задачах из репози-

тория UCI показали, что метод МПР увеличивает обобщающую способность решающих деревьев, по сравнению с методом МЭР. Улучшение наблюдалось при использовании как метода Монте-Карло, так и УНС-аппроксимации. Однако вычисление УНС-аппроксимации на порядки быстрее.

В **заключении** перечисляются основные результаты, выносимые на защиту.

Публикации по теме диссертации

1. *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Докл. всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 7–10.
2. *Ботов П. В.* Точные оценки вероятности переобучения для несимметричных многомерных семейств алгоритмов // Докл. межд. конф. Интеллектуализация обработки информации, ИОИ-8. — М.: МАКС Пресс, 2010. — С. 20–23.
3. *Botov P. V.* Exact estimates of the probability of overfitting for multidimensional modeling families of algorithms // ISSN 1054-6618, **Pattern Recognition and Image Analysis**, 2010, Vol. 20, No. 4, pp. 52–65. — Pleiades Publishing, Ltd., 2010.
4. *Ботов П. В.* Уменьшение вероятности переобучения итерационных методов статистического обучения // Докл. всеросс. конф. Математические методы распознавания образов-15. — М.: МАКС Пресс, 2011. — С. 44–47.