На правах рукописи

ИВАХНЕНКО АНДРЕЙ АЛЕКСАНДРОВИЧ

КОМБИНАТОРНЫЕ ОЦЕНКИ ВЕРОЯТНОСТИ ПЕРЕОБУЧЕНИЯ И ИХ ПРИМЕНЕНИЕ В ЛОГИЧЕСКИХ АЛГОРИТМАХ КЛАССИФИКАЦИИ

01.01.09 — дискретная математика и математическая кибернетика

Автореферат диссертации на соискание ученой степени кандидата физико-математических наук

Москва — 2010

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, профессор, **Каркищенко Александр Николаевич**
	кандидат физико-математических наук, **Чехович Юрий Викторович**
Ведущая организация:	Московский государственный университет им. М. В. Ломоносова, факультет ВМК

Защита диссертации состоится « ____ » _____ 2010 г. в _____ час. на заседании диссертационного совета Д 212.156.05 при Московском физико-техническом институте (государственном университете) по адресу: 141700, г. Долгопрудный Московской обл., Институтский пер. д.9, ауд. 903 КПМ.

С диссертацией можно ознакомиться в библиотеке МФТИ (ГУ).

Автореферат разослан « ____ » _____ 2010 г.

Учёный секретарь диссертационного совета
Д 212.156.05 О. С. Федько

Общая характеристика работы

Диссертационная работа относится к математической теории распознавания и классификации и посвящена проблеме повышения обобщающей способности логических алгоритмов классификации, основанных на поиске информативных конъюнктивных закономерностей в массивах прецедентных данных с вещественными, порядковыми и номинальными признаками.

Актуальность темы. Логические алгоритмы классификации широко используются для автоматизации принятия решений в трудноформализуемых областях, таких как медицинская диагностика, геологическое прогнозирование, кредитный скоринг, направленный маркетинг и т. д. Их преимуществом является возможность содержательной интерпретации как внутреннего строения алгоритма, так и принимаемых им решений на естественном языке в терминах предметной области. Однако качество классификации (обобщающая способность) логических алгоритмов, как правило, немного уступает более сложным конструкциям, таким как бустинг или бэггинг над решающими деревьями, которые являются «чёрными ящиками» и не обладают свойством интерпретируемости. Поэтому актуальной задачей является повышение обобщающей способности логических алгоритмов классификации без существенного усложнения их внутреннего строения.

Стандартные подходы теории статистического обучения дают слишком осторожные, пессимистичные оценки обобщающей способности. Эффекты переобучения, возникающие при поиске логических закономерностей, являются относительно слабыми. Они существенно зависят от конкретной выборки данных и потому плохо описываются стандартными оценками. Недав-

но разработанная комбинаторная теория переобучения даёт существенно более точные оценки вероятности переобучения. Актуальность данной диссертационной работы связана ещё и с тем, что она является первым примером практического применения комбинаторной теории переобучения.

Цель работы — получение комбинаторных оценок обобщающей способности логических закономерностей и разработка на их основе новых методов повышения качества логических алгоритмов классификации.

Научная новизна. Получена новая комбинаторная оценка вероятности переобучения, зависящая от характеристик расслоения и связности семейства алгоритмов. Комбинаторные оценки, полученные ранее другими авторами, либо были существенно менее точными, либо опирались на сильные дополнительные предположения о существовании в семействе корректного или хотя бы единственного лучшего алгоритма, что почти невозможно гарантировать на практике.

Впервые описана структура классов эквивалентности конъюнктивных логических закономерностей при разнотипных исходных данных, включающих количественные, порядковые и номинальные признаки.

Предложен новый метод вычисления поправки на переобучение, позволяющий улучшить широкий класс стандартных критериев информативности логических закономерностей.

Методы исследования. Для получения оценок вероятности переобучения использована слабая (перестановочная) вероятностная аксиоматика, комбинаторная теория переобучения, элементы комбинаторки, теории вероятностей и теории графов. Для проверки точности комбинаторных оценок проведены вы-

числительные эксперименты на модельных данных.

Для практического применения полученных оценок вероятности переобучения достаточно встроить процедуру вычисления поправок на переобучение в любой стандартный критерий информативности. Других модификаций не требуется, что позволяет усовершенствовать широкий класс эвристических методов построения логических алгоритмов классификации. Проведены эксперименты на реальных задачах классификации, подтвердившие, что указанная модификация приводит к повышению обобщающей способности алгоритмов классификации.

Положения, выносимые на защиту.

1. Общая комбинаторная оценка вероятности переобучения, зависящая от характеристик расслоения и связности семейства алгоритмов.

2. Описание структуры классов эквивалентности семейства пороговых конъюнкций над количественными, порядковыми и номинальными признаками.

3. Эффективный метод вычисления поправок на переобучение в критериях информативности для широкого класса эвристических алгоритмов поиска логических закономерностей.

4. Экспериментальное подтверждение того, что предложенный метод приводит к повышению обобщающей способности алгоритмов классификации, представляющих собой композиции логических закономерностей.

Теоретическая значимость. Данная работа вносит существенный вклад в развитие комбинаторной теории переобуче-

ния и является первым успешным примером практического применения комбинаторных оценок вероятности переобучения.

Практическая значимость. Предложенные методы расширяют область применимости логических алгоритмов классификации и повышают качество решения задач классификации в тех предметных областях, где применение логических алгоритмов продиктовано соображениями интерпретируемости.

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

- Всероссийская конференция «Математические методы распознавания образов» ММРО-12, 2005 г. [1];
- Международная конференция «Интеллектуализация обработки информации» ИОИ-6, 2006 г. [2];
- 49-я научная конференция МФТИ, 2006 г. [3];
- Всероссийская конференция «Математические методы распознавания образов»ММРО-13, 2007 г. [4, 5, 6];
- Международная конференция «Интеллектуализация обработки информации» ИОИ-7, 2008 г. [7];
- Всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [8].
- 52-я научная конференция МФТИ, 2009 г. [9];
- Международная конференция «Интеллектуализация обработки информации», ИОИ-8, 2010 г. [12, 13].

- Научные семинары отдела Интеллектуальных систем Вычислительного центра РАН и кафедры «Интеллектуальные системы» МФТИ, 2006 – 2010 г.г.

Публикации по теме диссертации. Всего публикаций по теме диссертации — 13, в том числе в изданиях из Списка, рекомендованного ВАК РФ — одна [11].

Структура и объём работы. Работа состоит из введения, четырёх глав, заключения, списка использованных источников, включающего 58 наименований. Общий объём работы составляет 100 страниц.

Краткое содержание работы по главам

Глава 1. Проблема обобщающей способности в логических алгоритмах классификации

1.1. Задача классификации. Пусть имеется пространство *объектов* $\mathbb{X}$ и конечное множество *классов* $\mathbb{Y}$. Объекты описываются набором n *признаков* $f_j\colon \mathbb{X} \to D_j,\ j = 1, \ldots, n$, где D_j — множество допустимых значений j-го признака. Таким образом, произвольному объекту $x \in \mathbb{X}$ соответствует *признаковое описание* $\big(f_1(x), \ldots, f_n(x)\big)$. В зависимости от множества D_j признаки делятся на типы: бинарные ($D_j = \{0, 1\}$), номинальные (D_j — конечное множество), порядковые (D_j — конечное упорядоченное множество), количественные ($D_j = \mathbb{R}$). *Целевой признак* $y^*\colon \mathbb{X} \to \mathbb{Y}$ известен только на объектах *обучающей выборки* $X = \{x_i\}_{i=1}^{\ell} \subset \mathbb{X},\ \ y_i = y^*(x_i)$. Требуется построить *алгоритм классификации* — функцию $a\colon \mathbb{X} \to \mathbb{Y}$, которая по признаковому описанию произвольного объекта x предсказывала бы его

класс $y^*(x)$, то есть приближала бы неизвестную функцию y^* на всём множестве $\mathbb{X}$.

Методом обучения называется отображение $\mu\colon (\mathbb{X}\times\mathbb{Y})^\ell \to A$, которое произвольной конечной выборке $X \subset \mathbb{X}$ с известными классификациями ставит в соответствие алгоритм $a = \mu X$ из заданного семейства алгоритмов классификации A.

Задана бинарная функция $I(a,x)$, называемая *индикатором ошибки*. Если $I(a,x) = 1$, то говорят, что алгоритм a ошибается на объекте x. Обычно полагается $I(a,x) = [a(x) \neq y^*(x)]$.

Обозначим через $n(a,X) = \sum\limits_{x\in X} I(a,x)$ число ошибок, а через $\nu(a,X) = n(a,X)/|X|$ — частоту ошибок алгоритма a на выборке X. Метод обучения μ называется *минимизацией эмпирического риска* (МЭР), если

$$\mu X \in A(X) = \operatorname*{Arg\,min}_{a\in A} n(a,X),$$

и *пессимистичным* методом МЭР, если

$$\mu X = \operatorname*{arg\,max}_{a\in A(X)} n(a,\mathbb{X}).$$

1.2. Проблема переобучения. Если алгоритм a доставляет минимум функционалу $\nu(a,X)$ на заданной обучающей выборке X, то это ещё не гарантирует, что он будет хорошо приближать целевую зависимость на произвольной контрольной выборке $\bar{X} = (x'_i, y'_i)_{i=1}^k$. Когда качество работы алгоритма на новых объектах, не вошедших в состав обучения, оказывается существенно хуже, чем на обучающей выборке, говорят, что имеет место *переобучение* или переподгонка (overfitting).

В комбинаторной теории переобучения предполагается, что $\mathbb{X}$ — конечное множество из $L = \ell + k$ объектов, и все его разби-

ения на обучающую выборку длины ℓ и контрольную длины k равновероятны. Основной задачей является получение как можно более точной верхней оценки *вероятности переобучения*

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \eta(\varepsilon) = \mathsf{P}\big[\nu(\mu X, \bar{X}) - \nu(\mu X, X) > \varepsilon\big]. \tag{1.1}$$

Коль скоро такая оценка получена, можно утверждать, что $\nu(\mu X, \bar{X}) \leqslant \nu(\mu X, X) + \varepsilon(\eta)$ с вероятностью $1 - \eta$, достаточно близкой к единице, где $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon)$.

1.3. Логические алгоритмы классификации представляют собой композиции информативных логических правил.

Логическое правило — это функция вида $r\colon \mathbb{X} \to \{0, 1\}$ из некоторого параметрического семейства функций R. Обычно к правилам предъявляют требование *интерпретируемости*: правило должно зависеть от небольшого числа признаков и допускать запись на естественном языке. В данной работе рассматривается наиболее распространённый тип правил — конъюнкции пороговых предикатов:

$$r(x) \equiv r(x; c^1, \ldots, c^n) = \prod_{j \in \omega} \big[x^j \lessgtr_j c^j\big], \tag{1.2}$$

где $x = (x^1, \ldots, x^n) \in \mathbb{R}^n$ — признаковое описание объекта x, $\omega \subseteq \{1, \ldots, n\}$ — подмножество признаков, $c^j \in D_j$ — порог по j-му признаку, $\lessgtr_j$ — одна из операций сравнения: $\{\leqslant, \geqslant\}$ для количественных и порядковых признаков, $=$ для номинальных.

Говорят, что правило r *выделяет* объект x, если $r(x) = 1$.

Логическая закономерность — это правило, удовлетворяющее требованию *информативности* — оно должно выделять достаточно много объектов одного из классов $y \in \mathbb{Y}$, и практически не выделять объекты других классов. Для поиска правил

класса y по обучающей выборке $X \subset \mathbb{X}$ в семействе $r \in R$ решается задача двухкритериальной оптимизации:

$$P_y(r, X) = \frac{1}{|X|} \sum_{x_i \in X} r(x_i) \, [y_i = y] \to \max;$$
$$N_y(r, X) = \frac{1}{|X|} \sum_{x_i \in X} r(x_i) \, [y_i \neq y] \to \min;$$

На практике двухкритериальную задачу сводят к однокритериальной и оптимизируют критерий информативности $H(P, N)$. Известны десятки различных критериев: энтропийный, индекс Джини, точный тест Фишера, тест χ^2, тест ω^2 и другие. Однако ни один из них нельзя назвать безусловно предпочтительным. Выбор критерия информативности является эвристикой.

Поскольку каждая закономерность выделяет лишь часть выборки, алгоритм классификации строится как композиция закономерностей. Одной из наиболее распространённых конструкций является *взвешенное голосование* закономерностей:

$$a(x) = \arg\max_{y \in Y} \sum_{r \in R_y} w_r r(x),$$

где R_y — множество правил класса y, $w_r \geqslant 0$ — вес правила r.

1.4. Постановка задачи. Чтобы алгоритм классификации не был переобучен, составляющие его закономерности также не должны быть переобучены. Поэтому первая задача, которая ставится в данной работе — получить оценки вероятности переобучения для семейства правил вида (1.2). Вторая задача — применить полученные оценки для улучшения стандартных критериев информативности.

Глава 2. Комбинаторные оценки вероятности переобучения для пороговых конъюнкций

2.1. Оценки расслоения–связности. Метод порождающих и запрещающих множеств, предложенный К. В. Воронцовым, основан на гипотезе, что для любого алгоритма $a \in A$ можно записать необходимое и достаточное условие того, что он будет выбран методом обучения μ по выборке X. В данной работе предлагается более простое необходимое условие, которое приводит к верхней оценке вероятности переобучения.

Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ множество A состоит из алгоритмов a с попарно различными *векторами ошибок* $\big(I(a, x_i)\big)_{i=1}^{L}$.

Гипотеза 2.1. Множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого алгоритма $a \in A$ можно указать два множества: *порождающее* $X_a \subset \mathbb{X}$ и *запрещающее* $X'_a \subset \mathbb{X}$ такие, что

$$\big[\mu X{=}a\big] \leqslant \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big]. \tag{2.1}$$

Введём гипергеметрическое распределение $h_L^{\ell,m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ и его функцию распределения $H_L^{\ell,m}(z) = \sum\limits_{s=0}^{\lfloor z \rfloor} h_L^{\ell,m}(s)$.

Теорема 2.1. *Если гипотеза 2.1 верна, то для любого $\varepsilon \in (0,1)$ справедлива верхняя оценка вероятности переобучения*

$$Q_\varepsilon \leqslant \sum_{a \in A} P_a H_{L_a}^{\ell_a, m_a}\big(s_a(\varepsilon)\big); \quad \mathsf{P}[\mu X{=}a] \leqslant P_a = \frac{C_{L_a}^{\ell_a}}{C_L^\ell}, \tag{2.2}$$

где $L_a = L - |X_a| - |X'_a|$, $\ell_a = \ell - |X_a|$, $m_a = n(a, \mathbb{X}\backslash X_a \backslash X'_a)$, $s_a(\varepsilon) = \frac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_a)$.

Определим отношение порядка на алгоритмах $a \leqslant b$ как

естественное отношение порядка на их векторах ошибок: $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$. Определим метрику на алгоритмах как хэммингово расстояние между их векторами ошибок: $\rho(a,b) = \sum\limits_{i=1}^{L} |I(a,x) - I(b,x)|$. Алгоритмы a и b называются *связанными*, если $a \leqslant b$ и $\rho(a,b) = 1$.

Лемма 2.2. *Если μ — метод пессимистичной минимизации эмпирического риска, то* (2.1) *выполнено, если положить*

$$X_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon I(a,x) < I(b,x),\ a \leqslant b,\ \rho(a,b) = 1\big\},$$
$$X'_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon I(b,x) < I(a,x),\ b \leqslant a\big\}.$$

Величину $q(a) = |X_a|$ будем называть *связностью*, а величину $r(a) = |X'_a|$ — *подчинённостью* алгоритма a.

Теорема 2.3. *Если μ — метод пессимистичной минимизации эмпирического риска, то для вероятности переобучения справедлива оценка сверху:*

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \sum_{a \in A} \frac{C^{\ell-q(a)}_{L-q(a)-r(a)}}{C^\ell_L} H^{\ell-q(a),\, n(a,\mathbb{X})-r(a)}_{L-q(a)-r(a)} \left(\tfrac{\ell}{L}(n(a,\mathbb{X}) - \varepsilon k)\right).$$

Если в этой оценке положить $q(a) = r(a) = 0$ для всех a, то получится оценка, известная в теории Вапника-Червоненкиса:

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \sum_{a \in A} H^{\ell, n(a,\mathbb{X})}_L \left(\tfrac{\ell}{L}(n(a,\mathbb{X}) - \varepsilon k)\right),$$

частично учитывающая расслоение семейства алгоритмов по числу ошибок $n(a,\mathbb{X})$, и совсем не учитывающая связность.

2.2. Структура классов эквивалентности семейства пороговых конъюнкций. Оценки вероятности переобучения,

полученные выше для алгоритмов классификации, легко переносятся на логические правила, если соответствующим образом определить индикатор ошибки: $I(r, x_i) = \big[r(x_i) \neq [y_i{=}y]\big]$.

Допустим, что значения x_i^j каждого признака $j \in \omega$ попарно различны на объектах $x_i \in \mathbb{X}$. Тогда без ограничения общности можно предполагать, что все признаки принимают целые значения $1, \ldots, L$, и никакие два объекта не имеют равных значений одного признака. Значения порогов c^j в правилах (1.2) также имеет смысл выбирать только из целых значений $0, \ldots, L$.

Пусть u и v — произвольные объекты из $\mathbb{X}$. Будем говорить, что объект u *доминируется* множеством объектов $S \subseteq \mathbb{X}$ (запись: $u \prec S$), если для каждого $j \in \omega$ существует объект $s \in S$, такой, что $u^j < s^j$.

Определение 2.1. *Подмножество $S \subseteq \mathbb{X}$ называется недоминирующимся, если любой объект $s \in S$ не доминируется подмножеством $S \backslash s$.*

Введём множество всех недоминирующихся подмножеств мощности q, $M_q = \big\{S \subseteq \mathbb{X}\colon |S| = q,\ \forall s \in S\ \ s \not\prec S\backslash s\big\}$.

Лемма 2.4. *Для любого объекта x из недоминирующегося подмножества S найдется хотя бы один признак $j \in \omega$, для которого $x^j = \max\limits_{s \in S} s^j$, причём*

$$\bigcup_{j=1}^{n} \operatorname{Arg}\max_{s \in S}(s^j) = S.$$

Поставим в соответствие подмножеству S правило

$$r(x, S) = r(x; \max_{x \in S} x^1, \ldots, \max_{x \in S} x^n).$$

Правила эквивалентны, если их векторы ошибок совпадают.

Лемма 2.5. *Пусть $E \subset R$ — класс эквивалентности правил, $c^j(r)$ — j-й параметр правила r. Классу E принадлежит правило*

$$r_E(x) = r\big(x; \min_{r' \in E} c^1(r'), \ldots, \min_{r' \in E} c^n(r')\big).$$

Будем называть правило $r_E(x)$ *стандартным представителем* класса эквивалентности E.

Теорема 2.6. *Существует взаимно однозначное соответствие между множеством всех классов эквивалентности и множеством всех недоминирующихся подмножеств.*

Для каждого класса эквивалентности E существует единственное недоминирующееся подмножество S: $r_E(x) = r(x, S)$. Число классов эквивалентности равно $\sum\limits_{q=0}^{n} |M_q|$.

Полученные результаты обобщаются на случай, когда признак может принимать одинаковые значения на различных объектах, а также на порядковые и номинальные признаки.

2.3. Эффективные методы перебора классов эквивалентности и вычисления оценок вероятности переобучения. Для эффективного вычисления оценки вероятности переобучения по Теореме 2.3 предлагается послойный алгоритм перебора классов эквивалентности. Вклады слоёв в вероятность переобучения быстро убывают с ростом номера слоя, равного числу ошибок правила $n(r, \mathbb{X})$. Поэтому для вычисления оценки достаточно обойти лишь несколько нижних слоёв, содержащих небольшую долю всех классов эквивалентности, и тем самым избежать перебора основной массы классов эквивалентности.

2.4. Информативность пороговых конъюнкций с поправкой на переобучение. Стандартные критерии информативно-

сти строятся исключительно по обучающей выборке и не учитывают эффект переобучения. При фиксированном подмножестве признаков ω переобучение возникает вследствие оптимизации порогов c_j, $j \in \omega$. Оно проявляется в том, что значения критериев $P(r, X)$ и $N(r, X)$ на обучающей выборке X оказываются оптимистично смещёнными относительно $P(r, \bar{X})$ и $N(r, \bar{X})$ на контрольной выборке $\bar{X}$. Предположим, что оценка $P(r, X)$ завышена на ΔP, оценка $N(r, X)$ занижена на ΔN. Предлагается с помощью оценок вероятности переобучения, и затем обращения этих оценок, оценить величину смещений ΔP, ΔN, чтобы при отборе закономерностей пользоваться модифицированным критерием информативности $H(P - \Delta P, N + \Delta N)$. Величина смещений ΔP, ΔN нетривиальным образом зависит от выборки, поэтому введение таких поправок может существенно влиять на отбор закономерностей во множества R_y.

2.5. Эксперименты на модельных данных. Генерируются две модельные выборки длины $L = 200$ с $n = 2$ признаками. В первой выборке существует единственная наилучшая закономерность и эффект расслоения проявляется максимально четко. Вторая выборка несколько зашумлена, в ней также существует наилучшая закономерность, но эффект расслоения выражен значительно слабее. Выборки подобраны таким образом, чтобы при обучении по случайным подвыборкам длины $\ell = 100$ информативности найденных в обеих выборках закономерностей в среднем совпадали. На контрольных подвыборках длины $k = L - \ell$, информативность найденной закономерности в первой подвыборке оказывается заметно выше чем во второй. Таким образом, метод оптимизации информативности сильнее переобучается на второй выборке. Однако стандартные критерии информативности не позволяют выявить эти различия. После

введения поправки описанным выше методом модифицированный критерий информативности начинает надёжно отличать более информативную выборку от менее информативной.

Глава 3. Методы построения логических алгоритмов классификации

В данной главе рассматриваются эвристические методы поиска логических закономерностей и построения логических алгоритмов классификации как композиций закономерностей, реализованные автором в рамках библиотеки Forecsys LogicPro. Для всех методов приводятся их описания в виде псевдокода, показывающего ключевые идеи программной реализации, но скрывающего второстепенные технические детали.

3.1. Методы оценивания информативности правил. Рассматриваются критерии информативности, реализованные в библиотеке LogicPro: ε, δ-критерий, гипергеометрический критерий (точный тест Фишера), энтропийный, индекс Джини.

3.2. Методы поиска информативных конъюнкций. Рассматриваются следующие методы: случайный поиск, случайный поиск с адаптацией, усечённый поиск в ширину, построение Парето-оптимального фронта и Парето-расслоения по двум критериям (P, N). Кроме того, рассматривается алгоритм бинаризации исходных данных, применяемый на подготовительной стадии перед поиском конъюнкций, а также методы стабилизации и редукции, применяемые после поиска конъюнкций для улучшения их качества, удаления переобученных и дублирующих конъюнкций.

3.3. Методы построения композиций информативных правил. Рассматриваются следующие методы построения алгоритмов классификации: жадный алгоритм покрытий для построения решающего списка (комитета старшинства), алгоритм бустинга для построения взвешенного голосования, алгоритм построения простого или взвешенного голосования.

3.4. Методы оценивания апостериорных вероятностей. Во многих практических задачах требуется, чтобы алгоритм классификации не только относил объект к тому или иному классу, но и выдавал оценки апостериорных вероятностей принадлежности объекта классу. Для взвешенного голосования эта задача решается с помощью калибровки Платта, однако когда объект покрывается малым числом закономерностей, точность данной оценки может оказаться невысокой. В случае решающего списка данный метод вообще неприменим. В обоих случаях проблема решается путём оценивания апостериорной вероятности для каждого правила в отдельности.

Пусть $r_y(x)$ — закономерность класса y. Оценку апостериорной вероятности того, что объект x принадлежит классу $c \in Y$ при условии, что он выделяется закономерностью r_y, в данной работе предлагается вычислять по формуле

$$P\big(c \mid r_y(x)=1\big) = \frac{P_c(r_y) - \Delta P_c}{(P_c(r_y) - \Delta P_c) + (N_c(r_y) + \Delta N_c)},$$

где ΔP_c и ΔN_c — поправки на переобучение, вычисляемые по описанной выше методике. Без этих поправок оценка была бы смещённой (оптимистично завышенной в случае $c = y$ или пессимистично заниженной при $c \neq y$).

Глава 4. Вычислительные эксперименты на реальных данных

В данной главе описываются вычислительные эксперименты на реальных данных из репозитория UCI, в которых сравниваются различные эвристические алгоритмы построения логических алгоритмов классификации. Показывается, что введение поправок на переобучение практически во всех задачах и для всех критериев информативности приводит к улучшению обобщающей способности отдельных закономерностей, и, как следствие, алгоритма классификации в целом.

В **Заключении** приведены основные результаты работы.

Основные результаты диссертации:

1. Получена общая комбинаторная оценка вероятности переобучения, зависящая от характеристик расслоения и связности семейства алгоритмов.

2. Описана структура классов эквивалентности семейства пороговых конъюнкций над количественными, порядковыми и номинальными признаками.

3. Предложен эффективный метод вычисления поправок на переобучение в критериях информативности для широкого класса эвристических алгоритмов поиска логических закономерностей.

4. Экспериментально показано, что предложенный метод приводит к повышению обобщающей способности алгоритмов классификации, представляющих собой композиции логических закономерностей.

Публикации по теме диссертации

1. *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Система кредитного скоринга на основе логических алгоритмов классификации // Докл. всеросс. конф. Математические методы распознавания образов-12.— М.: МАКС Пресс, 2005.— С. 349–353.
2. *Воронцов К. В., Ивахненко А. А.* Эмпирические оценки локальной функции роста в задачах поиска логических закономерностей // Искусственный Интеллект.— Донецк, 2006.— С. 281–284.
3. *Ивахненко А. А.* Обобщающая способность логических закономерностей // Труды 49-й научной конференции МФТИ. Часть VII.— 2006.— С. 286–287.
4. *Ивахненко А. А., Воронцов К. В.* Верхние оценки переобученности и профили разнообразия логических закономерностей // Докл. всеросс. конф. Математические методы распознавания образов-13.— М.: МАКС Пресс, 2007.— С. 33–37.
5. *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Применение логических алгоритмов классификации в задачах кредитного скоринга и управления риском кредитного портфеля банка // Докл. всеросс. конф. Математические методы распознавания образов-13.— М.: МАКС Пресс, 2007.— С. 484–488.
6. *Кузнецов М. Р., Туркин П. Ю., Воронцов К. В., Дьяконов А. Г., Ивахненко А. А., Сиваченко Е. А.* Прогнозирование результатов хирургического лечения атеросклероза на основе анализа клинических и иммунологических данных // Докл. всеросс. конф. Математические методы распознава-

ния образов-13. — М.: МАКС Пресс, 2007. — С. 537–539.

7. *Воронцов К. В., Ивахненко А. А.* Мета-обучение критериев информативности логических закономерностей // Докл. межд. конф. Интеллектуализация обработки информации, ИОИ-7. — Симферополь, 2008. — С. 52–54.
8. *Воронцов К. В., Ивахненко А. А., Инякин А. С., Лисица А. В., Минаев П. Ю.* «Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации // Докл. всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 503–506.
9. *Ивахненко А. А.* Верхняя оценка вероятности переобучения логических закономерностей // Труды 52-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». — 2009. — С. 64–67.
10. *Ивахненко А. А.* Точная верхняя оценка вероятности переобучения для корректных логических правил // В сборнике «Модели и методы обработки информации». — М.: МФТИ., 2009. — С. 62–64.
11. *Ивахненко А. А.* Точная верхняя оценка вероятности переобучения для корректных логических правил // **Труды МФТИ**. — Том 2 — № 3 — М.: МФТИ., 2010. — С. 81–87.
12. *Ивахненко А. А.* О вероятности переобучения пороговых конъюнкций // Докл. межд. конф. Интеллектуализация обработки информации, ИОИ-8. — М.: МАКС Пресс, 2010. — С. 57–60.
13. *Лисица А.В., Воронцов К.В., Ивахненко А.А., Инякин А.С., Синцова В.В.* Системы тестирования алгоритмов машинного обучения: MLcomp, TunedIt и Полигон // Докл.

межд. конф. Интеллектуализация обработки информации, ИОИ-8. — М.: МАКС Пресс, 2010. — С. 157–160.

В работах с соавторами лично соискателем сделано следующее:

- разработана и реализована библиотека логических алгоритмов Forecsys LogicPro [1, 5, 6];
- реализована система тестирования алгоритмов классификации и расчета статистик [8, 13];
- проведены эксперименты по исследованию обобщающей способности логических алгоритмов классификации, построению профиля разнообразия закономерностей [2, 4];
- проведены эксперименты по изучению зависимости обобщающей способности логических закономерностей от их параметров, таких как длинна закономерности, ранг по информативности и пр. [7]

ИВАХНЕНКО АНДРЕЙ АЛЕКСАНДРОВИЧ

КОМБИНАТОРНЫЕ ОЦЕНКИ ВЕРОЯТНОСТИ ПЕРЕОБУЧЕНИЯ И ИХ ПРИМЕНЕНИЕ В ЛОГИЧЕСКИХ АЛГОРИТМАХ КЛАССИФИКАЦИИ

Автореферат

Подписано в печать: 15.11.2010

Заказ № 4522 Тираж – 100 экз.

Печать трафаретная. Объем: 1 усл.п.л.

Типография «11-й ФОРМАТ»

ИНН 7726330900

115230, Москва, Варшавское ш., 36

(499) 788-78-56

www.autoreferat.ru