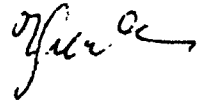


На правах рукописи



Чистяков Сергей Павлович

**МЕТОД МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА
ПРИ ИНДУКТИВНОМ ПОСТРОЕНИИ БАЗ ЗНАНИЙ**

05.13.18 — математическое моделирование, численные
методы и комплексы программ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Петрозаводск 2006

Работа выполнена в Институте прикладных математических исследований Карельского научного центра РАН.

Научный руководитель:

доктор физико-математических наук, профессор Павлов Ю.Л.

Официальные оппоненты:

доктор технических наук, доцент Рогов А.А.

кандидат физико-математических наук, с.н.с. Полин А.К.

Ведущая организация:

Пермский государственный университет

Защита состоится "3" марта 2006г. в 16 часов на заседании диссертационного совета Д 212.190.03 в Петрозаводском государственном университете по адресу: 185910, г. Петрозаводск, пр. Ленина, 33.

С диссертацией можно ознакомиться в библиотеке Петрозаводского государственного университета.

Автореферат разослан "30" января 2006г.

Ученый секретарь
диссертационного совета
Д 212.190.03



Поляков В.В.

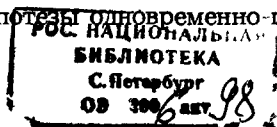
2006 А
2426

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Ключевой проблемой теории и практики экспертных систем (ЭС) является проблема построения баз знаний (БЗ). На современном этапе развития технологии ЭС широкое распространение получила парадигма построения БЗ непосредственно из данных. Процесс приобретения знаний непосредственно из данных носит название *индуктивного обучения*. Достоинства такого подхода состоят в том, что он позволяет в кратчайшие сроки и с минимальными затратами осуществлять построение БЗ. Поэтому разработка методов и алгоритмов индуктивного построения БЗ в настоящее время представляет собой чрезвычайно актуальную задачу как в теоретическом, так и в прикладном плане.

В диссертации рассматривается две проблемы. Первая состоит в индуктивном построении БЗ, использующих модель представления знаний в виде систем правил с весами, которые могут рассматриваться как некоторый классификатор. Основное внимание было уделено разработке и модификации схем реализации и алгоритмов метода структурной минимизации эмпирического риска (МСМЭР), позволяющих избежать перепогонки (*overfitting*) — явления, хорошо известного в статистике и состоящего в том, что вероятностно-статистическая модель, построенная по некоторой выборке, фактически описывает только саму эту выборку и непригодна в качестве модели всей рассматриваемой генеральной совокупности. Актуальность этой проблематики в задаче индуктивного построения БЗ определяется тем, что перепогонка приводит к снижению качества классификатора, индуцированного построенной БЗ.

Второй круг вопросов связан с построением БЗ статистических ЭС. В диссертации рассматривается задача использования априорной информации для оптимального выбора уровней значимости статистических критериев, совместно применяемых для проверки одной и той же нулевой гипотезы против нескольких классов альтернатив. Актуальность этой задачи состоит в том, что возможный характер отклонений от нулевой гипотезы известен не всегда и, вследствие этого, не всегда можно из нескольких критериев проверки одной и той же нулевой гипотезы против различных классов альтернатив выбрать критерий, наиболее подходящий в конкретной ситуации. Поэтому для повышения достоверности статистических выводов часто приходится осуществлять проверку нулевой гипотезы одновременно по несколь-



ким статистическим критериям, что расширяет область альтернатив и позволяет минимизировать количество ошибок второго рода.

Цель исследования. Целью диссертации являлась разработка методов и алгоритмов индуктивного построения БЗ на основе МСМЭР, позволяющих избежать эффекта переподгонки и учитывающих априорную информацию, а также разработка методов и алгоритмов построения наилучшего совместного критерия при наличии информации о возможных альтернативах.

Объекты исследования. Объектом исследований были индуктивные алгоритмы построения БЗ, схемы реализации МСМЭР, методы и алгоритмы нахождения и упорядочения множества наиболее информативных признаков, статистические критерии проверки гипотез и методы назначения априорных распределений.

Методы исследования. Основными методами исследований являлись вероятностные и статистические методы, в частности методы байесовской и робастной статистики. При сравнении качества алгоритмов дискретизации и алгоритмов поиска и упорядочения множества наиболее информативных признаков использовалась методика эмпирического сравнения соответствующих индуктивных алгоритмов на реальных задачах.

Научная новизна. В диссертации предложены новые схемы реализации МСМЭР, учитывающие частичную информацию об оценках вероятности ошибочной классификации во всех классах решающих функций, предложен новый метод дискретизации непрерывных признаков, основанный на использовании критерия однородности χ^2 с ограничением на минимальное число значений признака в интервалах дискретизации. Разработаны новые методы учета априорной информации при совместном применении статистических критериев.

Основные положения диссертации, выносимые на зачет. На зачет выносятся:

1. Схемы реализации МСМЭР и соответствующие алгоритмы, использующие частичную информацию о качестве классификаторов, индуцированных последовательностью минимальных БЗ.
2. Алгоритмы предварительной обработки обучающей выборки, ориентированные на индуктивное построение БЗ.
3. Методы и алгоритмы построения наилучшего совместного кри-

терия при наличии априорной, частичной априорной или эмпирической информации о возможных альтернативах.

4. Комплекс программ индуктивного построения БЗ и результаты его применения при решении задач кардиологии и социологии.

Связь работы с крупными научными программами, темами. Результаты диссертации получены в рамках двух научно-исследовательских тем Института прикладных математических исследований Карельского научного центра РАН: "Исследование и разработка методов создания интеллектуальных систем статистического анализа данных" (№гос. регистрации 01.9.80009162) и "Разработка методов исследования случайных структур и их применения при принятии статистических решений" (№гос. регистрации 01.200.202223). В 1997–1998 годах работа по теме диссертации была поддержана грантом РФФИ №97-01-00554 "Моделирование процесса принятия решений при выборе методов статистического анализа".

Апробация результатов диссертации. Основные результаты работы докладывались на Шестой научной конференции стран СНГ "Применение многомерного статистического анализа в экономике и оценке качества продукции" (Москва, 1997г.), Юбилейной научной конференции Карельского научного центра Российской академии наук посвященной 275-летию РАН (Петрозаводск, 1999г.), Пятой Петрозаводской международной конференции "Вероятностные методы в дискретной математике" (Петрозаводск, 2001г.), четвертом Всероссийском симпозиуме по прикладной и промышленной математике (Петрозаводск, 2003г.) и на Ученом совете Института прикладных математических исследований Карельского научного центра РАН.

Публикация результатов. Основные результаты диссертации опубликованы в двенадцати работах, из них: одна статья в центральном журнале, две статьи в сборниках трудов международных конференций, шесть статей в сборниках трудов Института прикладных математических исследований КарНЦ РАН и Петрозаводского государственного университета.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, списка литературы, заключения и четырех приложений. Общий объем диссертации составляет 142 страницы. Список литературы содержит 87 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во введении дан краткий обзор задачи индуктивного обучения, приведено обоснование актуальности темы диссертации, сформулированы ее цели и основные положения, выносимые на защиту.

Первая глава носит вспомогательный характер. В ней приведены определения основных понятий и описаны методы и алгоритмы, используемые в последующих главах: статистические решающие функции (р.ф.), метод минимизации эмпирического риска (ММЭР) и МСМЭР, кросс-проверка, дается определение БЗ, используемое в диссертации, описаны статистические критерии, алгоритмы дискретизации и упорядочения признаков.

Пусть $\mathbf{X} = (X_1, X_2, \dots, X_n)$ — случайный вектор с множеством возможных значений X и с распределением $P(\mathbf{x})$. Предполагается, что существует "учитель", который классифицирует значения $\mathbf{X}$, т.е. относит их к одному из k классов, помеченных метками из множества $D = \{0, 1, \dots, k-1\}$ в соответствии со значением случайной величины (с.в.) Y , имеющей распределение $P(y|\mathbf{x})$, $y \in D$. Координаты вектора $\mathbf{X}$ называются *входными признаками*, а с.в. Y — *классовым признаком*. Распределения $P(\mathbf{x})$ и $P(y|\mathbf{x})$ неизвестны, однако известно, что они существуют, т.е. существует совместное распределение $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$. Пусть также определено некоторое множество р.ф. $\mathcal{F} = \{f_\alpha(\mathbf{x}) : \mathbf{x} \in X, \alpha \in \Lambda\}$, то есть при любом $\alpha \in \Lambda$ функция $f_\alpha(\mathbf{x}) : X \rightarrow D$. Качество $R(\alpha)$ р.ф. $f_\alpha(\mathbf{x})$ или *риск* определяется как вероятность различной классификации вектора $\mathbf{x}$ учителем и с помощью р.ф. $f_\alpha(\mathbf{x})$, а именно $R(\alpha) = P\{Y \neq f_\alpha(\mathbf{X})\}$. Эта величина носит также название *вероятности ошибочной классификации* (в.о.к.). Обозначим $\alpha_0 = \arg \min_{\alpha \in \Lambda} R(\alpha)$ и предположим, что имеется случайная выборка $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_l, y_l)\}$ из распределения $P(\mathbf{x}, y)$ объема l , называемая *обучающей выборкой*. Задача нахождения р.ф. $f_{\alpha_0}(\mathbf{x})$, минимизирующей риск $R(\alpha)$ по обучающей выборке D для произвольного распределения $P(\mathbf{x}, y)$, называется задачей *классификации с учителем*. Решающие функции $f_\alpha(\mathbf{x})$ в задаче классификации с учителем называются также *классификаторами*, а отображение, сопоставляющее каждой обучающей выборке некоторый классификатор, *индуктивным алгоритмом*. Одним из методов приближенного нахождения $f_{\alpha_0}(\mathbf{x})$ является ММЭР. Обозначим

$$\chi_\alpha(\mathbf{x}, y) = \begin{cases} 0, & y = f_\alpha(\mathbf{x}) \\ 1, & y \neq f_\alpha(\mathbf{x}) \end{cases}.$$

Заменяем риск $R(\alpha)$ его оценкой (эмпирическим риском):

$$R_{emp}(\alpha) = R_{emp}(\alpha, \mathcal{D}) = \frac{1}{l} \sum_{i=1}^l \chi_{\alpha}(x_i, y_i). \quad (1)$$

Суть ММЭР состоит в том, что в качестве р.ф., приближающей $f_{\alpha_0}(x)$, выбирается функция $f_{\alpha^*}(x)$, где

$$\alpha^* = \arg \min_{\alpha \in \Lambda} R_{emp}(\alpha).$$

Развитием ММЭР является МСМЭР. Пусть имеется набор классов р.ф. (к.р.ф.) $\mathcal{F}_i = \{f_{\alpha}(x) : x \in X, \alpha \in \Lambda_i\}$, $i = 0, 1, \dots, N$, таких, что $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_N$ и Ψ — некоторый функционал, определенный на множестве р.ф. $\mathcal{F}_N$, т.е. $\Psi : \mathcal{F}_N \rightarrow R^1$. Пусть $f_{\alpha_i^*}(x)$ — р.ф., минимизирующая эмпирический риск в классе $\mathcal{F}_i$. Тогда

$$i_0 = \arg \min_{i=0,1,\dots,N} \Psi(f_{\alpha_i^*}) \quad (2)$$

определяет номер к.р.ф., оптимальный с точки зрения МСМЭР, а р.ф. $f_{\alpha_{i_0}^*}(x)$ представляет собой оптимальный классификатор по МСМЭР. Функционал Ψ , рассматриваемый в диссертации, сопоставляет р.ф. $f_{\alpha}(x)$ статистическую оценку в.о.к. $R(\alpha)$, получаемую при помощи метода блочной кросс-проверки.

В диссертации рассматривается частный случай задачи классификации с учителем, в котором входные признаки $X_1, X_2, \dots, X_n$ являются дискретными с.в., принимающими конечное число значений. Упорядоченную пару "признак — значение" будем называть *категорией* и обозначать Xx (признак X принял значение x). Категории классового признака Y обозначаются $C_i^* = Yi$, $i = 0, 1, \dots, k-1$. Через C^* будем обозначать произвольную категорию классового признака. *Комбинацией* $C = X_{\alpha_1}x_{\alpha_1\beta_1} \& X_{\alpha_2}x_{\alpha_2\beta_2} \& \dots \& X_{\alpha_r}x_{\alpha_r\beta_r}$, где $\alpha_i = 1, 2, \dots, n$, $\beta_i = 1, 2, \dots, r_i$, $i = 1, 2, \dots, r$ называется конъюнкция категорий, не содержащая двух одинаковых признаков, т.е. $X_{\alpha_i} \neq X_{\alpha_j}$, при $i \neq j$. *Длиной комбинации* $length(C) = r$ назовем число включенных в нее категорий. Комбинация C_1 называется *подкомбинацией* комбинации C_2 (обозначение $C_1 \subset C_2$), если каждая категория из C_1 также включена и в C_2 . Введем также пустую комбинацию C^0 с нулевой длиной. По определению $D(C^0) = D$ и $C^0 \subset C$ для

любой комбинации C . Упорядоченную пару комбинации C и некоторой категории классового признака C^* назовем *эмпирической импликацией* и будем обозначать $C \Rightarrow C^*$. Эмпирическую импликацию с приписанным ей весом $w \in (0, 1)$ будем называть *правилом* и обозначать $C \Rightarrow C^* \langle w \rangle$. Левую часть правила C будем называть *антецедентом*, а правую C^* — *консеквентом* правила. Под *априорными правилами* понимаются правила вида $C^0 \Rightarrow C_i^* \langle w_i^0 \rangle$. Множество априорных правил будем обозначать $\mathcal{R}_0$. Для двух правил $C_1 \Rightarrow C^* \langle w_1 \rangle$ и $C_2 \Rightarrow C^* \langle w_2 \rangle$ с одинаковым консеквентом C^* определена *операция композиции* весов $\oplus$:

$$w_1 \oplus w_2 = \frac{w_1 w_2}{w_1 w_2 + (1 - w_1)(1 - w_2)}.$$

Пусть $\mathcal{R}$ некоторое множество (система) правил, содержащее все априорные правила, т.е. $\mathcal{R}_0 \subseteq \mathcal{R}$. Для произвольной комбинации C и категории C^* определим *композиционный вес* $W(C^* | C, \mathcal{R})$ как результат применения операции композиции $\oplus$ к весам всех правил из $\mathcal{R}$ вида $C' \Rightarrow C^* \langle w \rangle$, где $C' \subset C$. Обозначим через Δ некоторое фиксированное подмножество множества всех импликаций. Импликации из множества Δ будут называться *допустимыми*.

Определение 1. Пусть имеется обучающая выборка $\mathcal{D}$, оператор композиции $\oplus$, статистический критерий K , множество допустимых импликаций Δ и классовый признак Y . Множество правил $KB = KB(\mathcal{D}, \oplus, K, \Delta, Y)$ будем называть (*классификационной*) базой знаний (по отношению к $\mathcal{D}, \oplus, K, \Delta, Y$), если

1. KB содержит все априорные правила, т.е. $\mathcal{R}_0 \subseteq KB$;
2. для любой комбинации C входных признаков такой, что импликация $C \Rightarrow C^*$ допустима, нулевая гипотеза $H_0 : P(C^* | C) = W(C^* | C, KB)$ не отвергается критерием K против соответствующей двусторонней альтернативы.

Определение 2. База знаний KB будет называться минимальной, если она есть подмножество каждого множества правил, являющегося базой знаний по отношению к $\mathcal{D}, \oplus, K, \Delta, Y$.

Пусть KB — некоторая БЗ. Определим р.ф. (классификатор) $f_{KB} : X \rightarrow D$ следующим образом:

$$f_{KB}(x) = \arg \max_i W(C_i^* | C(x), KB),$$

где $\mathbf{x} = (x_1, x_2, \dots, x_n) \in X$ и комбинация $C(\mathbf{x})$ определяется соотношением $C(\mathbf{x}) = X_1 x_1 \& X_2 x_2 \& \dots \& X_n x_n$. Функцию $f_{KB}(\mathbf{x})$ будем называть р.ф., *индуцированной KB*.

Во второй главе приведены схемы реализации МСМЭР и соответствующие алгоритмы, использующие частичную информацию о качестве классификаторов, индуцированных последовательностью минимальных БЗ.

В первом разделе описаны индуктивные алгоритмы построения последовательности минимальных БЗ, использующие способы генерации комбинаций по возрастанию их длины и количества входящих в них наиболее информативных признаков. Обозначим $\Omega_1(L)$, $L = 0, 1, \dots, n$ — множество систем правил с длиной антецедента, не превосходящей L и $\mathcal{F}_1(L)$ — к.р.ф., индуцированный $\Omega_1(L)$. Легко видеть, что $\Omega_1(0) \subset \Omega_1(1) \subset \dots \subset \Omega_1(n)$ и $\mathcal{F}_1(0) \subset \mathcal{F}_1(1) \subset \dots \subset \mathcal{F}_1(n)$. Обозначим через $\Delta_1(L)$, $L = 0, 1, \dots, n$ — множество допустимых импликаций с длиной антецедента не более L . В Следующая теорема устанавливает свойства индуктивного алгоритма построения БЗ, использующего способ генерации комбинаций по возрастанию их длины (величина L_{max} является параметром алгоритмов и ее смысл понятен из формулировок нижеприведенных теорем).

Теорема 1. *1. каждая комбинация с допустимой частотой генерируется алгоритмом, причем только один раз;*

2. к моменту генерации первой комбинации длины $k = 1, 2, \dots, L_{max}$ все комбинации меньшей длины уже были сгенерированы;

3. для каждого $L = 0, 1, \dots, L_{max}$ множество правил БЗ $KB(L) \subset KB(L_{max}) = KB$ с длиной антецедента не более L является минимальной базой знаний по отношению к $D, \oplus, K, \Delta_1(L), Y$.

Пусть признаки $X_1, X_2, \dots, X_n$ расположены в порядке уменьшения априорной вероятности того, что данный признак понадобится при классификации, $\Omega_2(L)$, $L = 0, 1, \dots, n$ — множество систем правил, антецеденты которых содержат не более L первых признаков из данного упорядоченного списка и $\mathcal{F}_2(L)$ — к.р.ф., индуцированный $\Omega_2(L)$. Легко видеть, что $\Omega_2(0) \subset \Omega_2(1) \subset \dots \subset \Omega_2(n)$

и $\mathcal{F}_2(0) \subset \mathcal{F}_2(1) \subset \dots \subset \mathcal{F}_2(n)$. Обозначим $\Delta_2(L)$, $L = 0, 1, \dots, n$ — множество допустимых импликаций с количеством наиболее информативных признаков antecedента не более L . Следующая теорема устанавливает свойства индуктивного алгоритма построения БЗ, использующего способ генерации комбинаций по возрастанию количества входящих в них наиболее информативных признаков.

Теорема 2. *1. каждая комбинация с допустимой частотой генерируется алгоритмом, причем только один раз;*

2. к моменту генерации первой комбинации, содержащей некоторую категорию признака X_L , все комбинации, содержащие только категории признаков $X_1, X_2, \dots, X_{L-1}$, уже были сгенерированы;

3. для каждого $L = 0, 1, \dots, L_{max}$ множество правил $KB(L) \subset KB$, состоящее из правил, antecedенты которых содержат только категории признаков $X_1, X_2, \dots, X_L$, является минимальной базой знаний по отношению к $\mathcal{D}, \oplus, \mathbf{K}, \Delta_2(L), Y$.

Во втором разделе приведены схемы реализации МСМЭР, основанные на использовании частичной информации о качестве оптимальных классификаторов во всех к.р.ф. и дающие возможность осуществления автоматического или интерактивного управления процессом вычислений.

Пусть $P_J(L)$ — оценка в.о.к. оптимального классификатора $f_{\alpha_L^*}(\mathbf{x})$ в классе $\mathcal{F}_L$, $L = 0, 1, \dots, n$, полученная по J блокам кросс-проверки $J = 1, 2, \dots, M$. Определим величины

$$P_J^{\leq}(L) = \min_{1 \leq K \leq L} P_J(K) \quad \text{и} \quad P_J^{>}(L) = \min_{L < K \leq N} P_J(K).$$

Схема, предназначенная для автоматического управления поиском оптимального к.р.ф. включает следующие два этапа:

1. получение оценок в.о.к. оптимальных классификаторов во всех к.р.ф. по заданному числу блоков $J_0 < M$;
2. оставшиеся $M - J_0$ блоков обрабатываются по последовательной схеме начиная с класса $\mathcal{F}_0$ со следующим критерием прекращения

ния поиска: *остановить вычисления, как только*

$$\frac{P_M(K)}{P_M^{\leq}(K)} > \alpha \quad \text{и} \quad \frac{P_{J_0}^>(K)}{P_M^{\leq}(K)} > \beta.$$

Величины J_0 , $\alpha > 1$ и $\beta > 1$ являются параметрами критерия остановки процесса вычислений. Предложенная схема, помимо учета информации об оценках в.о.к. оптимальных классификаторов в K к.р.ф., полученных после обработки всех M блоков кросс-проверки учитывает и частичную информацию об оценках в.о.к. оптимальных классификаторов в $N-K$ к.р.ф., полученных после обработки $J_0 < M$ блоков кросс-проверки, что повышает надежность решений о прекращении вычислений за счет некоторого увеличения времени.

Более гибкие возможности управления представляет схема реализации МСМЭР, предназначенная для интерактивного управления поиском оптимального к.р.ф. Она предполагает, что каждый блок обрабатывается последовательно по возрастанию номеров к.р.ф. и происходит параллельный пересчет оценок в.о.к., так что в любой момент график текущих оценок в.о.к. визуально доступен для лица, осуществляющего управление ходом вычислений. Управляющее воздействие состоит в том, что происходит отказ от дальнейшей оценки в.о.к. в блоках $\mathcal{F}_K, \mathcal{F}_{K+1}, \dots, \mathcal{F}_n$, $K \leq n$, причем такое воздействие может осуществляться несколько раз в ходе процесса вычислений.

В разделе приведены алгоритмы, реализующие описанные выше схемы реализации МСМЭР и позволяющие одновременно осуществлять выбор оптимального уровня значимости критерия, используемого для проверки импликаций на статистическую значимость. Приведены результаты тестирования алгоритмов на реальных задачах, показывающие, что предложенные схемы позволяют значительно сократить время вычислений.

В третьей главе описаны алгоритмы *предварительной обработки* (применение которых предполагается до процесса индуктивного обучения) и результаты их тестирования на реальных задачах (для проверки на статистическую значимость использовался односторонний критерий знаков). Алгоритмы учитывают особенности индуктивных алгоритмов главы 2. Приведено описание прототипа программного комплекса СТАТКОП для индуктивного построения БЗ и результаты его применения при решении задач кардиологии и социологии.

В первом разделе описан алгоритм нахождения и упорядочения множества наиболее информативных признаков.

Обозначим $KB(X_{\alpha_1}, X_{\alpha_2}, \dots, X_{\alpha_r})$, $\alpha_i \in \{1, 2, \dots, n\}$, $r \leq n$ минимальную БЗ, содержащую правила, антецедент которых включает только категории входных признаков $X_{\alpha_1}, X_{\alpha_2}, \dots, X_{\alpha_r}$. В разработанном алгоритме поиска и упорядочения множества наиболее информативных признаков в качестве меры информативности набора признаков $X_{\alpha_1}, X_{\alpha_2}, \dots, X_{\alpha_r}$ (учитывая то, что в основе рассматриваемых в диссертации алгоритмов лежит ММЭР) используется эмпирический риск $R_{emp}(X_{\alpha_1}, X_{\alpha_2}, \dots, X_{\alpha_r})$ классификатора, индуцированного соответствующей минимальной БЗ. Результаты тестирования алгоритма на реальных задачах, показывают, что применение алгоритма в сочетании с индуктивными алгоритмами главы 2 позволяет:

1. значимо (при уровне значимости 0.05) уменьшить количество отобранных информативных признаков по сравнению с алгоритмом упорядочения признаков ReliefF;
2. значимо (при уровне значимости 0.05) уменьшить количество правил в БЗ по сравнению с подходом, не использующим отбор наиболее информативных признаков.

Во втором разделе приведен алгоритм ChiSplit, предназначенный для дискретизации непрерывных признаков. Для сравнительной оценки качества дискретизации, определяемой различными разбиениями области значений непрерывного признака в обучающей выборке, в алгоритме использована статистика критерия однородности χ^2 . Особенностью алгоритма является введение параметра, задающего минимальное число значений непрерывного дискретизируемого признака, попадаемых в каждый интервал разбиения. Поскольку алгоритм ориентирован в качестве предварительной обработки для индуктивных алгоритмов, приведенных в главе 2, то введение этого параметра является естественным, так как комбинации с частотой ниже заданной в этих алгоритмах не анализируются.

Приведены результаты тестирования алгоритма на реальных задачах, показывающие, что применение алгоритма ChiSplit на стадии предварительной обработки статистически значимо (при уровне значимости 0.05) уменьшает в.о.к. соответствующих классификаторов по сравнению с алгоритмами ChiMerge, CAIM, TSE, KEX.

Во третьем разделе приводится краткое описание комплекса программ индуктивного построения БЗ "СТАТИстический КОНструктор Правил" (СТАТКОП) и результаты его применения при решении задач кардиологии и социологии.

Программное ядро СТАТКОП основано на алгоритмах, приведенных в главах 2 и 3 диссертации. Прототип является функционально полным, т.е. в нем реализованы все модули необходимые для его нормального функционирования, как вспомогательного характера (ввод и контроль обучающей выборки, сохранение и ввод ранее построенных БЗ для классификации), предварительной обработки (нахождение и упорядочение множества наиболее информативных признаков, дискретизация непрерывных признаков), модули обеспечивающие реализацию основной задачи (построение БЗ) и модули, обеспечивающие использование построенных БЗ для классификации.

В области кардиологии СТАТКОП использовался для определения факторов риска ранней госпитальной кардиальной летальности и построения классификаторов, позволяющих прогнозировать исход операции. Исходными данными для анализа послужили истории болезней 116 пациентов, которым выполнялась одномоментная коррекция клапанного порока и аортокоронарное шунтирование с 1991 по 1996 гг. в Научном центре сердечно-сосудистой хирургии им. А.Н. Бакулева РАМН. Обучающая выборка содержала 16 входных признаков, описывающих как предоперационное состояние пациентов, так и некоторые характеристики проведенных им операций (вид коррекции клапанов, количество шунтов и др.), а в качестве классового признака фигурировал исход операции (хороший или летальный). Для повышения достоверности выводов наряду с построением БЗ были применены и другие статистические методы, а именно, анализ таблиц сопряженности признаков и построение деревьев решений. Отметим, что классификаторы, индуцированные построенной БЗ и деревом решений основывались на одних и тех же трех признаках — предыдущие операции на сердце, кальциноз и коррекция клапанов (одноклапанная или многоклапанная). Содержательный анализ результатов применения всех упомянутых методов позволил отнести к наиболее важным факторам риска для рассматриваемой операции следующие: предыдущие операции на сердце, кальциноз, вид коррекции клапанов, тип порока и количество шунтов.

В области социологии СТАТКОП был использован при решении

задачи установления взаимосвязей между современным потенциалом сельских домохозяйств Карелии и их проективными стратегиями. Исходными данными для анализа послужили результаты выборочного обследования 186 сельских домохозяйств Республики Карелия, проведенного в 2004 году социологами Карельского научного центра РАН. Потенциал домохозяйств выражался через несколько групп структурных признаков: демографических, социальных, производственно-хозяйственных, бюджетных, потребительских, занятости. Структурные признаки в обучающей выборке играли роль входных признаков (их количество составило 87). В качестве классowego использовался признак, характеризующий предполагаемые намерения (проективные стратегии) домохозяйств и отражающий адаптационный процесс на уровне социальных ориентаций и предпочтений. Анализ и интерпретация правил БЗ, построенной с помощью СТАТКОП, позволил выявить наиболее характерные черты потенциала групп домохозяйств с одним и тем же типом проективной стратегии ("самодостаточные", "активные", "активно-пассивные", "пассивные", "эксклюзивные").

В четвертой главе рассматривается задача совместного применения статистических критериев. Пусть $K_1, K_2, \dots, K_m$ — статистические критерии проверки нулевой гипотезы H_0 против альтернатив $H_1, H_2, \dots, H_m$ соответственно. Обозначим через K критерий проверки нулевой гипотезы H_0 против "объединенной" альтернативы $H = H_1 \cup H_2 \cup \dots \cup H_m$ который отвергает гипотезу H_0 , если хотя бы один из критериев $K_1, K_2, \dots, K_m$ отвергает H_0 . Критерии такого вида будем называть *совместными*. Для уровня значимости β совместного критерия K выполнены неравенства $\max(\alpha_1, \alpha_2, \dots, \alpha_m) \leq \beta \leq \alpha_1 + \alpha_2 + \dots + \alpha_m$, где $\alpha_1, \alpha_2, \dots, \alpha_m$ — уровни значимости критериев $K_1, K_2, \dots, K_m$ соответственно. Выбор $\alpha_1, \alpha_2, \dots, \alpha_m$, являющийся предметом рассмотрения данной главы, в значительной степени определяется наличием (или отсутствием) у исследователя априорной информации о вероятностях альтернатив того или иного класса.

Пусть X — случайная величина (с.в.) с функцией распределения $F(x)$, принадлежащей классу распределений $F = \{F(x, \theta), \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$. Обозначим $\mathbf{X} = (X_1, X_2, \dots, X_n)$ случайную выборку из распределения с.в. X . Множество значений случайного вектора $\mathbf{X}$ будем обозначать X . Предположим, что каждое альтернативное распределение определяется вектором $\theta \in \Theta_H \subset \Theta$, и пусть на Θ_H существует априорное распределение Π . Пусть вектор $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$

и K_α означает соответствующий совместный критерий. Обозначим $\Lambda_\gamma = \{\alpha : \alpha_1 + \alpha_2 + \dots + \alpha_m = \gamma\}$. Определим риск $R(\Pi, \alpha)$ критерия K_α следующим образом:

$$R(\Pi, \alpha) = \int_{\Theta_H} P_\theta(X \notin K_\alpha) \Pi(d\theta),$$

где K_α — критическая область критерия K_α , определяемая соотношением

$$K_\alpha = \{x \in X : x \in \bigcup_{j=1}^m K_{\alpha_j}, \alpha \in \Lambda_\gamma\},$$

а K_{α_j} — критическая область критерия K_j , $j = 1, 2, \dots, m$. Совместный критерий K_{α_0} будем называть наилучшим совместным критерием (н.с.к.), если

$$\alpha_0 = \arg \min_{\alpha \in \Lambda_\gamma} R(\Pi, \alpha).$$

Предположим, что альтернативные распределения случайно генерируются в соответствии с (неизвестным) априорным распределением Π . Пусть $\theta_1, \theta_2, \dots, \theta_l \in \Theta_H$ — случайная выборка из распределения Π и $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = 1, 2, \dots, l$ — случайная выборка из распределения $F(x, \theta_i)$. Обозначим $T_j(X)$ статистику критерия K_j , $F_j(x)$ функцию распределения с.в. $T_j(X)$ в случае истинности нулевой гипотезы H_0 .

Теорема 3. 1. Существует с.в. Y с множеством возможных значений $\{0, 1\}$, множество $p.f.$

$$\mathcal{F} = \{f_\alpha(v) : v \in D^m, \alpha \in \Lambda_\gamma\},$$

определенных на единичном кубе D^m со значениями в $\{0, 1\}$ и распределение $W(dvdy)$ на $D^m \times \{0, 1\}$ такие, что

$$R(\Pi, \alpha) = \int_{\Theta_H} P_\theta(X \notin K_\alpha) \Pi(d\theta) = \int_{D^m \times \{0, 1\}} (y - f_\alpha(v))^2 W(dvdy).$$

2. Пусть критические области K_{α_j} критериев K_j , $j = 1, 2, \dots, m$ имеют вид $K_{\alpha_j} = \{x : T_j(x) \geq c_{\alpha_j}\}$, где $F_j(c_{\alpha_j}) = 1 - \alpha_j$. Тогда для любого $\epsilon > 0$

$$P\{|R(\Pi, \alpha_0) - R(\Pi, \alpha^*)| > \epsilon\} \rightarrow 0 \text{ при } l \rightarrow \infty,$$

где $\alpha^* = \arg \min_{\alpha \in \Lambda_\gamma} R_{emp}(\alpha)$ и

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^l (1 - f_\alpha(v_i))^2.$$

Во втором разделе рассматривается задача использования частичной априорной информации (ч.а.и.) для построения двусторонних статистических критериев. Пусть K_α — двусторонний критерий проверки простой нулевой гипотезы $H_0 : \theta = \theta_0$ против двусторонней альтернативы $H : \theta \neq \theta_0$, определяемый критической областью $K_\alpha = K_{\alpha_1}^- \cup K_{\alpha_2}^+$, где $\alpha = (\alpha_1, \alpha_2)$ и $K_{\alpha_1}^-$ и $K_{\alpha_2}^+$ критические области левостороннего и правостороннего критериев с уровнями значимости α_1 и α_2 соответственно. Пусть $q = (q_1, q_2, \dots, q_m)$ и $p = (p_1, p_2, \dots, p_{m-1})$ таковы, что $q_1 < q_2 < \dots < q_m$ и $\sum_{i=1}^{m-1} p_i = 1$. Обозначим

$$\Gamma = \{\Pi : \Pi(\theta \in \Theta_i) = p_i, i = 1, 2, \dots, m-1\}, \quad \Theta_i = (q_i, q_{i+1}], \quad (3)$$

$$R_\Gamma^+(\alpha) = \sup_{\Pi \in \Gamma} R(\Pi, \alpha), \quad R_\Gamma^-(\alpha) = \inf_{\Pi \in \Gamma} R(\Pi, \alpha).$$

Теорема 4. Пусть функция мощности $W(K_\alpha; \theta)$ критерия K_α монотонно убывает на интервале $(-\infty, \theta_{\min})$ и монотонно возрастает на интервале $(\theta_{\min}, +\infty)$. Тогда нижняя $R_\Gamma^-(\alpha)$ и верхняя $R_\Gamma^+(\alpha)$ границы риска критерия K_α определяются следующими соотношениями:

$$R_\Gamma^+(\alpha) = 1 - \sum_{i=1}^{m-1} p_i \inf_{\theta \in \Theta_i} W(K_\alpha; \theta), \quad R_\Gamma^-(\alpha) = 1 - \sum_{i=1}^{m-1} p_i \sup_{\theta \in \Theta_i} W(K_\alpha; \theta),$$

где

$$\inf_{\theta \in \Theta_i} W(K_\alpha; \theta) = \begin{cases} W(K_\alpha; q_{i+1}), & q_{i+1} < \theta_{\min} \\ W(K_\alpha; \theta_{\min}), & q_i < \theta_{\min} < q_{i+1} \\ W(K_\alpha; q_i), & q_i > \theta_{\min} \end{cases}$$

и

$$\sup_{\theta \in \Theta_i} W(K_\alpha; \theta) = \begin{cases} W(K_\alpha; q_i), & q_{i+1} < \theta_{\min} \\ \max\{W(K_\alpha; q_i), W(K_\alpha; q_{i+1})\}, & q_i < \theta_{\min} < q_{i+1} \\ W(K_\alpha; q_{i+1}), & q_i > \theta_{\min}. \end{cases}$$

Утверждение теоремы 4 использовано в алгоритме, разработанном для определения оптимальных значений уровней значимости левостороннего и правостороннего критериев на основе экспертной ч.а.и. о квантилях априорного распределения на множестве альтернатив, вводимой в алгоритм в интерактивном режиме.

В заключении сформулированы основные результаты диссертационной работы и направления дальнейших исследований.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Алекаян Б.Г., Скопин И.И., Никитина Т.Г., Капутин М.Ю., Захаров И.В., Стаферов А.В., Грошева Т.В., Чистяков С.П. Коронарная ангиопластика в этапном лечении больных приобретенными пороками сердца в сочетании с ИБС // Грудная и сердечно-сосудистая хирургия. – 2001. Вып. 6. – С. 73–76.
2. Белая Р. В., Морозова Т. В., Чистяков С. П. Проективные стратегии и современный потенциал сельских домохозяйств // Тр. ИПМИ КарНЦ РАН. – 2005. – Вып. 5. – С. 21–34.
3. Павлов Ю. Л., Харин В.Н., Чистяков С. П. Применение статистических методов для оценки точности решений экспертных систем // Тр. Петрозаводского гос. университета, сер. "Прикладная математика и информатика". – 1997. – Вып. 6. – С. 179–186.
4. Чистяков С.П. Об определении уровней значимости одновременно применяемых статистических критериев // Тр. ИПМИ КарНЦ РАН. – 1999. – Вып. 1. – С. 55–60.
5. Чистяков С.П. Применение метода структурной минимизации эмпирического риска при индуктивном построении баз знаний // Тр. ИПМИ КарНЦ РАН. – 2002. – Вып. 3. – С. 213–225.
6. Чистяков С.П. Об одном алгоритме дискретизации непрерывных признаков // Тр. ИПМИ КарНЦ РАН. – 2004. Вып. 4. – С. 203–212.
7. Чистяков С.П. О поиске множества наиболее информативных признаков при индуктивном построении баз знаний // Тр. ИПМИ КарНЦ РАН, Вып. 5, 2004. – С. 128–135.

8. Чистяков С.П. Об автоматизации построения баз знаний экспертных систем // Обзорение прикладной и промышленной математики, т. 8, вып. 1, 2001. – С. 375–376.
9. Чистяков С.П. Об использовании априорной информации для определения уровней значимости совместно применяемых статистических критериев // Обзорение прикладной и промышленной математики, Т. 7, 2001. – С. 152–153.
10. Чистяков С.П. Применение кросс-проверки для оптимизации выбора параметров индуктивных алгоритмов построения систем правил // Обзорение прикладной и промышленной математики. – 2003. Т. 10, вып. 1. – С. 252.
11. Chistjakov S.P. On using of partial prior information for statistical tests construction // Proceeding of the sixth international conference "Computer Data Analysis and Modelling: Robustness and computer intensive methods", Minsk, 2001. – P. 104–109.
12. Chistjakov S.P. On joint using of statistical tests // Proceeding of Fifth International Petrozavodsk conference on Probability methods in discrete mathematics. VSP, Utrecht, 2004. – P. 159–162.

Изд. лиц. № 00041 от 30.08.99. Подписано в печать 17.01.06.
Формат 60х84 $\frac{1}{16}$. Бумага офсетная. Гарнитура «Times».
Уч.-изд. л. 1,0. Усл. печ. л. 1,1. Тираж 100 экз. Изд. № 15. Заказ 561

Карельский научный центр РАН
Редакционно-издательский отдел
185003, Петрозаводск, пр. А. Невского, 50

2006 A
2426

- 2426