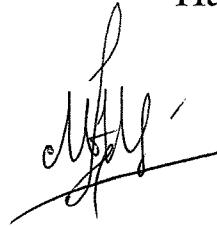


На правах рукописи



Шарков Максим Анатольевич

**СИНТЕЗ И АНАЛИЗ НЕПАРАМЕТРИЧЕСКИХ МОДЕЛЕЙ
СТОХАСТИЧЕСКИХ ЗАВИСИМОСТЕЙ И РАСПОЗНАВАНИЯ
ОБРАЗОВ В УСЛОВИЯХ МАЛЫХ ВЫБОРОК**

05.13.01 - Системный анализ, управление и обработка информации
(по отраслям информатика, вычислительная техника и управление)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Красноярск - 2007

Работа выполнена в Политехническом институте Сибирского федерального университета и Институте вычислительного моделирования СО РАН (г. Красноярск).

НАУЧНЫЙ РУКОВОДИТЕЛЬ:

доктор технических наук
профессор,
Заслуженный деятель науки РФ
Лапко Александр Васильевич

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ:

доктор технических наук,
профессор
Медведев Александр Васильевич
кандидат технических наук
Молоков Вячеслав Витальевич

ВЕДУЩАЯ ОРГАНИЗАЦИЯ:

Сибирский государственный
технологический университет

Защита состоится «12» октября 2007 года в 14:00 часов на заседании диссертационного совета Д.212.099.06 при ФГОУ ВПО «Сибирский федеральный университет» по адресу: 660074, г. Красноярск, ул.Киренского, 26. 660074, ауд. Д 501.

Факс: (3912) 43-06-92 (ПИ СФУ, для каф. АОИ)
E-mail: sovet@front.ru
Телефон: (3912) 91-22-36 (ПИ СФУ, каф. АОИ)

С диссертацией можно ознакомиться в библиотеке Политехнического института Сибирского федерального университета

Автореферат разослан «___» _____ 2007 г.

Учёный секретарь
диссертационного совета,

д.т.н

С.А. Бронов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Большинство статистических методов восстановления стохастических зависимостей и распознавания образов ориентировано на представительные обучающие выборки. Однако при решении прикладных задач часто располагают ограниченным объемом наблюдений – короткой либо малой выборкой.

Проблема анализа малых выборок является наименее исследованной в теории обучающихся систем. Для «обхода» проблем малых выборок широкое распространение получили методы, основанные на принципах декомпозиции систем и последовательные процедуры формирования решений, которые характеризуются недостаточной вычислительной эффективностью. В связи с этим, возникает потребность в разработке моделей восстановления стохастических зависимостей и распознавания образов в условиях малых выборок, обеспечивающих высокое качество и обоснованность получаемых результатов, а также невысокую ресурсоемкость при решении задач обработки информации.

Основные научные результаты диссертации получены в рамках планов научных исследований Института вычислительного моделирования СО РАН «Разработка гибридных интеллектуальных информационных технологий на основе интеграции геоинформационных, нейросетевых, экспертных и аналитических систем» ПСО №242 от 03.07.03 и Красноярского государственного технического университета, а также в соответствии с грантом Президента №МД-2130.2005.9.

Тема диссертации соответствует перечню «Критические технологии РФ» по направлению – компьютерное моделирование.

Народно-хозяйственная проблема. Нестационарность объекта исследования, высокая стоимость и сложность получения дополнительных данных обуславливает возникновение задач обработки информации в условиях малых выборок. Получаемые в этом случае решающие правила не всегда обеспечивают приемлемые для практики результаты, так как информация малых обучающих выборок затрудняет оценивание вероятностных характеристик изучаемых закономерностей. Данная проблема наиболее актуальна для социальных, медико-биологических и технических систем.

Объектом исследования являются методы обработки статистических выборок малого объема.

Предметом исследования являются непараметрические модели восстановление стохастических зависимостей и распознавание образов в условиях малых выборок

Цель научного исследования состоит в разработке методических, алгоритмических и информационных средств оценивания плотностей вероятности, синтеза алгоритмов распознавания образов и моделей восстановления стохастических зависимостей в условиях выборок малого объема, основанных на принципах имитации систем и методах непараметрической статистики.

Цель достигается путём решения следующих задач:

1. Разработка методики синтеза непараметрических моделей многомерных стохастических зависимостей и распознавания образов на основе принципов имитации систем, обеспечивающей эффективное использование информации обучающих выборок малого объема.
2. Обоснование предложенного направления исследований на основе анализа асимптотических свойств непараметрических моделей распознавания образов и восстановления стохастических зависимостей.
3. Разработка процедуры продолжения случайных последовательностей на основе анализа методики синтеза непараметрической оценки плотности вероятности в условиях малых выборок.
4. Создание информационных средств, реализующих непараметрические методы обработки информации в условиях малых выборок и их применение при исследовании социальных систем.

Методы исследования. Для решения поставленных задач использовались аппарат теории вероятности и непараметрической статистики, методы распознавания образов и восстановления многомерных стохастических зависимостей, принципы имитации систем.

Основные научные результаты:

1. Непараметрические оценки плотности вероятности, модели многомерных стохастических зависимостей и алгоритмы распознавания образов в условиях малых выборок, основанные на принципах имитации систем.
2. Количественная взаимосвязь между характеристиками обучающей выборки, параметрами имитационной процедуры формирования дополнительной статистической информации и непараметрических алгоритмов их обработки.
3. Методика продолжения случайных последовательностей с применением аппарата имитационного моделирования и методов непараметрической статистики.

Научная новизна. Впервые с позиций принципов имитации систем и методов непараметрической статистики теоретически обоснованы алгоритмы решения задач восстановления стохастических зависимостей, распознавания образов в условиях малых выборок.

Значение для теории. Результаты работы позволяют повысить эффективность построения моделей стохастических зависимостей и алгоритмов распознавания образов, а также открывают возможность построения непараметрических решающих правил в задачах классификации и моделирования неопределенных систем при обработке малых выборок.

Значение для практики. Разработаны информационные средства синтеза и анализа непараметрических моделей восстановления стохастических зависимостей и распознавания образов, ориентированные на исследование объектов различной природы в условиях малых выборок.

Критерии статистического оценивания условий преимущества предлагаемых моделей создают методическую и алгоритмическую основу автоматизации их проектирования при построении типовой информационной системы.

Созданы информационные средства, реализующие непараметрические методы обработки информации в условиях малых выборок, которые адаптированы для исследования динамики показателей преступности в регионах России.

Использование результатов диссертации. Разработанные методы, алгоритмы и информационные средства зарегистрированы в Отраслевом фонде алгоритмов и программ (свидетельство о регистрации № 6787) и используются для оценивания состояния преступности в регионе в учебном процессе Сибирского юридического института МВД РФ.

Личный вклад автора. Выбор направления исследований малых выборок выполнен автором совместно с научным руководителем. Все результаты получены лично автором.

Из шести публикаций пять подготовлены и опубликованы автором единолично.

Апробация работы. Основные положения диссертации представлялись и обсуждались на Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии» в г. Томске в марте 2006 года и марте 2007 года, на Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых «Молодежь и наука: начало XXI века» в мае 2005 года. Результаты работы докладывались на научных семинарах факультета информатики и процессов управления Красноярского политехнического института и Института вычислительного моделирования СО РАН.

Результаты исследований включались в научные отчеты Института вычислительного моделирования СО РАН, представлялись в отчетах гранта Президента РФ №МД-2130.2005.9.

Публикации. По результатам работы опубликовано 6 статей, в том числе в журнале «Вестник КрасГАУ», внесенном в перечень ведущих рецензируемых журнальных изданий.

Структура и объем работы. Диссертация состоит из введения, пяти глав, заключения, списка использованной литературы (131 наименование), содержит 124 страницы машинописного текста, иллюстрируется 29 рисунками.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность рассматриваемой в работе проблемы, определены цель и задачи исследования, выделены основные положения работы, имеющие научную новизну и практическую значимость.

В первой главе рассматриваются традиционные статистические методы обработки информации в условиях малых выборок, которые различаются принципами «обхода» возникающих проблем и процедурами реализации данных методов. На этой основе предложена методика классификации статистических моделей обработки информации, позволяющая систематизировать существующие подходы и определить новые постановки задач моделирования.

Проведен анализ основных методов обработки малых выборок, который позволил выявить ряд недостатков:

- формируемые с помощью бутстреп-метода частные выборки из исходной являются зависимыми, что создает значительные трудности при исследовании свойств строящихся статистических моделей;
- использование последовательной процедуры формирования решений в методе группового учета аргументов связаны со значительными вычислительными затратами;
- применение принципа декомпозиции исходной задачи в коллективах решающих правил в конечном счете приводит к снижению точности синтезируемых моделей.

Непараметрические методы обработки данных ориентируются на общие сведения об исследуемой системе, что обеспечивает возможность построения универсальных алгоритмов и моделей, не зависящих от природы изучаемых объектов. Однако, непараметрические методы применимы только к репрезентативным выборкам. Поэтому возникает актуальная проблема их развития в условиях малых выборок.

Во второй главе предложена методика увеличения объема исходных данных в задаче восстановления многомерной плотности вероятности в условиях малых выборок с использованием принципов имитационного моделирования.

Пусть имеется исходная выборка $V = (x^i, i = \overline{1, n})$ статистически независимых наблюдений случайной величины $x \in R^1$ с неизвестной плотностью вероятности $p(x)$. Объем данной выборки n считается малым и не позволяет обосновано применять методы непараметрической статистики. Используя методы имитационного моделирования, увеличим объем исходных данных. С этой целью в β -окрестности каждой i -ой ситуации выборки проведем m имитаций случайной величины x_Δ с законом распределения $p_\Delta(x_\Delta)$ и нулевым математическим ожиданием.

По результатам моделирования сформируем статистическую выборку $V_2 = (x^i + x_\Delta^j, i = \overline{1, n}, j = \overline{1, m})$ с законом распределения, соответствующим смеси плотностей вероятности

$$\bar{p}(x) = \frac{1}{n} \sum_{i=1}^n p_\Delta^i(x_\Delta). \quad (1)$$

Нетрудно заметить, что непараметрическая оценка (1) имеет вид

$$\bar{p}(x) = (nm)^{-1} \sum_{i=1}^n \sum_{j=1}^m \Phi\left(\frac{x - x^i - x_\Delta^j}{c}\right), \quad (2)$$

где $\Phi(u)$ -ядерная функция, удовлетворяющая условиям:

$$\Phi(u) \geq 0, \int \Phi(u) du = 1, \Phi(u) = \Phi(-u), \int u^\nu \Phi(u) du < \infty, \nu \geq 2,$$

$$\Phi(u_2) \geq \Phi(u_1), \forall |u_2| \leq |u_1|;$$

$c = c(n)$ - параметр размытости ядерной функции.

В многомерном случае $x \in R^k$ оценка плотности вероятности имеет вид

$$\bar{p}(x) = (nm)^{-1} \sum_{i=1}^n \sum_{j=1}^m \prod_{v=1}^k \frac{1}{c_v} \Phi\left(\frac{x_v - x_v^i - x_{v\Delta}^j}{c_v}\right). \quad (3)$$

Рекомендуемым критерием выбора параметра размытости $c(n)$ является максимум функция правдоподобия

$$L(c) = \prod_{k=1}^{n \times m} \bar{p}(x^k), \quad (4)$$

где, например, для статистики (3)

$$\bar{p}(x^k) = \frac{1}{(n \times m - 1)c} \sum_{\substack{i=1 \\ i \neq k}}^{n \times m} \Phi\left(\frac{x^k - x^i}{c}\right).$$

Доказана теорема о том, что для достаточно гладких плотностей вероятности существуют условия асимптотической несмешённости и состоятельности их непараметрических оценок типа (3). В частности, показано, что для равномерного закона распределения $p(x_\Delta)$ должны выполняться требования $c \rightarrow 0$, дисперсия случайной величины x_Δ $\sigma^2 \rightarrow 0$, а $nm c \rightarrow \infty$ при $n \rightarrow \infty$ и $m \rightarrow \infty$.

Доказательство асимптотической сходимости непараметрических статистик, предназначенных для обнаружения закономерностей в условиях малых выборок, позволяет аналитически обосновать методику их синтеза. На этой основе определить количественную взаимосвязь между характеристиками обучающей выборки, параметрами имитационной процедуры формирования дополнительной статистической информации и исследуемых алгоритмов их обработки, что необходимо для объяснения результатов вычислительных экспериментов.

Проведен анализ асимптотического выражения среднеквадратического отклонения $W(x, \bar{x})$ на всем диапазоне изменения переменных x, \bar{x} . Считая, что $p(\bar{x}) = (2\beta)^{-1} \forall x \in [-\beta; \beta]$ и пренебрегая в процессе преобразований величинами малости $\beta^3, \beta^2 c^2, c^4$ и $\beta^3 / (nmc)$, проинтегрировав выражения $W(x, \bar{x})$ по переменным x, \bar{x} , получим

$$W \sim \frac{\Delta}{2n\beta} + \frac{2\|p(x)\|^2 \beta}{m},$$

где Δ - длина интервала изменения x .

Как и следовало ожидать, с ростом m оценка среднеквадратического отклонения стремится к пределу $\frac{\Delta}{2n\beta}$. Причем, зависимость W от β при конкретных значениях m и n имеет экстремальный характер и при оптимальном

$$\beta^* = \left(\frac{m\Delta}{4n\|p(x)\|^2} \right)^{\frac{1}{2}} \quad (5)$$

достигает своего минимума.

Из анализа (5) следуют вполне очевидные соотношения между параметрами β^*, m и Δ . Интервал генерирования искусственной обучающей последовательности увеличивается с ростом области определения $p(x)$ и количества имитаций m , снижается по мере увеличения объема n исходной выборки.

Определено необходимое требование на количество имитаций m процедуры генерирования искусственной обучающей выборки

$$m > 2,56\Delta\|p(x)\|^2 \left(n^3 / \left((\|\Phi\|^2)^4 \|p_{(x)}^{(2)}\|^2 \right) \right)^{\frac{1}{5}},$$

при котором статистика (3) будет обладать более высокими аппроксимационными свойствами по сравнению с традиционной непараметрической оценкой плотности вероятности ядерного типа.

Полученные аналитические выводы подтверждают принципиальную возможность использования статистики типа (3) при обработке малых выборок. Асимптотические выражения для смещения и среднеквадратического отклонения отражают количественную взаимосвязь между параметрами процедуры имитации и исходными данными.

Кроме того, данный факт обосновывает использование алгоритма синтеза непараметрических оценок плотности вероятности (3) в качестве процедуры продолжения случайных последовательностей (x^i , $i = \overline{1, n}$).

1. На оси случайных чисел $\varepsilon \in [0;1]$ определим n равных отрезков и сопоставим их с элементами исходной выборки (x^i , $i = \overline{1, n}$). Примем значение параметра $i = 1$.

2. При помощи датчика случайных чисел сгенерируем число ε^i и по факту попадания его значения в i -й интервал определим опорную точку x_i .

3. В β -окрестности i -ой точки исходной выборки формируем $n+j$ -ое значение случайной величины x

$$x^{n+j} = x^i + 2(0.5 - \varepsilon^j)\beta,$$

где ε^j -случайная величина с равномерным законом распределения в области $[0;1]$.

4. Повторить этапы 2,3 m раз до необходимого объема ситуаций последовательности.

Асимптотическая несмещеннность и состоятельность непараметрических оценки плотности вероятности полученной выборки объемом $(n+m)$ следует из доказательства теоремы о существовании условий асимптотической несмешенности и состоятельности непараметрических оценок типа (3).

По данным вычислительных экспериментов обоснованы требования к параметрам имитационной процедуры увеличения объема выборки, необходимые для получения эффективных оценок плотностей вероятности в условиях малых выборок:

- Установлено, что β -окрестности процедуры имитационного моделирования должны иметь минимальные возможные области пересечения друг другом при условии максимального перекрывания области определения исходной выборки;
- достаточное количество имитаций в пределах β -окрестностей равно 5-7.

В третьей главе предлагаются и исследуются непараметрические алгоритмы распознавания образов в условиях малых выборок.

Пусть дана выборка $(x^i, \sigma(i), i = \overline{1, n})$, составленная из признаков x^i классифицируемых объектов с неизвестной плотностью $p(x)$ и «указаний учителя» $\sigma(i)$ о принадлежности ситуации x^i к тому либо иному классу Ω_t , $t = \overline{1, M}$.

Для упрощения выкладок, без существенной потери получаемых результатов, рассмотрим методику построения уравнения разделяющей

поверхности в условиях малых выборок на примере двуальтернативной задачи распознавания образов в пространстве непрерывных признаков. В этом случае решающее правило имеет вид

$$m(x) : \begin{cases} x \in \Omega_1, \text{ если } f_{12}(x) > 0, \\ x \in \Omega_2, \text{ если } f_{12}(x) \leq 0, \end{cases} \quad (6)$$

где $f_{12}(x)$ - байесово уравнение разделяющей поверхности между классами Ω_1, Ω_2 .

Для построения решающего правила воспользуемся оценкой плотности вероятности типа (3).

Пусть n_1 и n_2 - количество ситуаций обучающей выборки $(x^i, \sigma(i), i = \overline{1, n})$, принадлежащих первому и второму классам. Тогда непараметрическая оценка уравнения разделяющей поверхности, например, соответствующая критерию максимального правдоподобия, может быть представлена в виде

$$\bar{f}_{12}(x) = \left(nm \prod_{v=1}^k c_v \right)^{-1} \sum_{i=1}^n \sigma(i) \sum_{j=1}^m \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i - \bar{x}_v^j}{c_v}\right), \quad (7)$$

$$\text{где } \sigma(i) = \begin{cases} (n_1 / n), \text{ если } x \in \Omega_1, \\ -(n_2 / n), \text{ если } x \in \Omega_2. \end{cases}$$

При оптимизации решающего правила (7) сначала осуществляется выбор параметра β процедуры генерирования искусственной обучающей последовательности $(x^i + \bar{x}^j, \sigma(i, j), j = \overline{1, m}, i = \overline{1, n})$, где $\sigma(i, j)$ - указания о принадлежности ситуации $x^i + \bar{x}^j$ к одному из классов.

Обозначив β_t - окрестность точки x^i из t -го класса через $q_t^i(\beta)$, а их пересечения для двух ситуаций x^i, x^j через $q_t^{ij}(\beta_t) = q_t^i(\beta_t) \cap q_t^j(\beta_t)$. Введем естественное требование: β -окрестности выбираются для каждого класса и достаточно полно покрывают область определения соответствующей ему части обучающей выборки при условии минимального их пересечения.

Данное положение реализуется путем решения задач

$$\min_{\beta_t} \bigcap_{\substack{i, j \in I_t \\ i \neq j}} q_t^{ij}(\beta_t) \quad \forall x^i \in \bigcup_{\substack{i, j \in I_t \\ i \neq j}} q_t^{ij}(\beta_t), t = 1, 2$$

На втором этапе определяются параметры m и c_v непараметрической оценки уравнения разделяющей поверхности из условия минимума эмпирической оценки распознавания образов в режиме «скользящего экзамена».

Для повышения эффективности непараметрических алгоритмов распознавания образов в условиях малых выборок возможно использование принципов коллективного оценивания. Пусть $\tilde{m}_{12}^j(x), j = \overline{1, M}$ - непараметрические решающие правила для двуальтернативной задачи распознавания образов, которые построены по выборкам

$(x^i + \bar{x}^j, \sigma(i, j), j = \overline{1, m}, i = \overline{1, n})$, отличающимся случайными последовательностями, «расширяющими» исходную обучающую выборку, при одних и тех же значениях параметров имитации m и β .

Воспользуемся одним из известных подходов коллективного оценивания, например, методом «голосования» и построим решающее правило

$$\tilde{m}_{12}(x) : \begin{cases} x \in \Omega_1, \text{ если } \frac{M_1}{M} \geq \frac{M_2}{M} \\ x \in \Omega_2, \text{ если } \frac{M_2}{M} \geq \frac{M_1}{M}; \end{cases}, \quad (8)$$

где M_j , $j = 1, 2$ - число «решений», которые принимают члены коллектива о принадлежности объекта с набором признаков x в пользу j -го класса.

В многоальтернативной постановке задачи распознавания образов каждый член коллектива $\tilde{m}_{12}^j(x)$, $j = \overline{1, M}$ использует решающее правило типа (8). Окончательное вывод, например $x \in \Omega_t$, принимается, если частота решений членов коллектива в пользу t -го класса максимальное.

Для исследования эффективности предлагаемого алгоритма распознавания образов в условиях малых выборок в сравнении с хорошо зарекомендовавшим себя на практике традиционным непараметрическим алгоритмом распознавания образов, синтез которого осуществляется по исходной информации, решалась двухальтернативная задача распознавания образов в пространстве признаков $x_v, v = \overline{1, k}$. Априорные вероятности классов $P_1 = P_2 = 0.5$. Законы распределения признаков в области первого класса формировались в соответствии с датчиками случайных чисел

$$x_v = a + \varepsilon(b - a),$$

$$x_{v+1} = (x_v)^2 - 6x_v + 10 + \sigma_1 \left(\sum_{i=1}^{p_1} \varepsilon^i - 0.5p_1 \right) \frac{6}{\sqrt{3p_1}}, \quad v \in I_h,$$

где параметры распределений $a = 1.5$, $b = 4.5$, $p_1 = 5$; среднеквадратическое отклонение $\sigma_1 = 0.7$; $\varepsilon \in [0; 1]$ - случайная величина с равномерным законом распределения; $I_h = (1, 3, 5, \dots)$ - множество нечётных чисел меньших k .

Признаки второго класса генерировались с нормальным законом

$$x_v = m + \sigma_2 \left(\sum_{i=1}^{p_2} \varepsilon^i - 0.5p_2 \right) \frac{6}{\sqrt{3p_2}}, \quad v = \overline{1, k},$$

при $p_2 = 5$, $\sigma_2 = 0.7$, $m = 3$.

Случайные величины $\bar{x} = (\bar{x}_v, v = \overline{1, k})$, обеспечивающие расширение исходной обучающей выборки, формировались при помощи генерации в β -окрестности каждой i -ой точки исходной выборки m точек датчиком случайных чисел

$$\bar{x}_v^j = x_v^i + 2(0.5 - \varepsilon^j)\beta, v = \overline{1, k}; j = \overline{1, m}; i = \overline{1, n},$$

где ε^j -случайная величина с равномерным законом распределения в области $[0;1]$.

Формирование многомерных случайных величин основывалась на независимой генерации их координат.

Оценки вероятностей ошибок распознавания образов $\bar{\rho}$ при различных условиях классификации n, m, k и β вычислялись по контрольной выборке $n_k = 1000$, полученной с помощью приведённых выше датчиков случайных чисел.

Вычислительный эксперимент при фиксированных условиях исследований осуществлялся $N = 10$ раз и полученные результаты усреднялись.

Непараметрические алгоритмы распознавания образов в условиях малых выборок обеспечивают достоверное снижение ошибки распознавания образов на контрольных выборках по сравнению с традиционным классификатором ядерного типа, что особо проявляется при выборках существенно малого объема (рисунок 1).

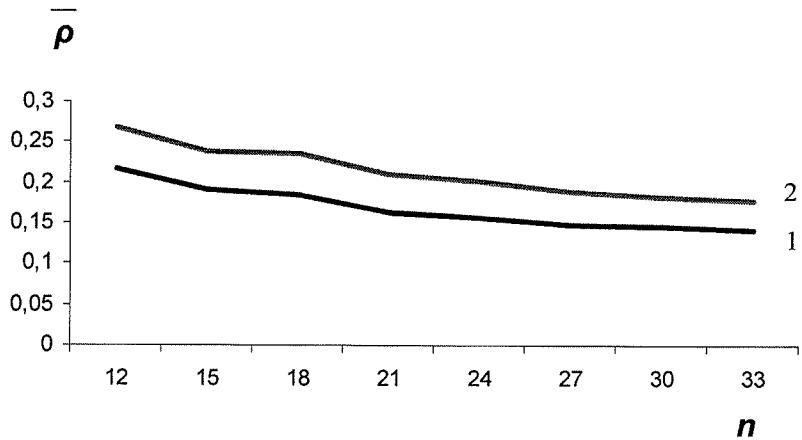


Рисунок 1 – Зависимость оценки вероятности ошибки $\bar{\rho}$ распознавания образов от объема n исходной обучающей выборки в условиях $\beta = 0.06$, $m = 7$ и $k = 4$. Кривые 1, 2 соответствуют исследуемому и традиционному непараметрическим алгоритмам распознавания образов.

С ростом размерности признаков k достоверное преимущество исследуемого алгоритма над традиционным непараметрическим классификатором сохраняется.

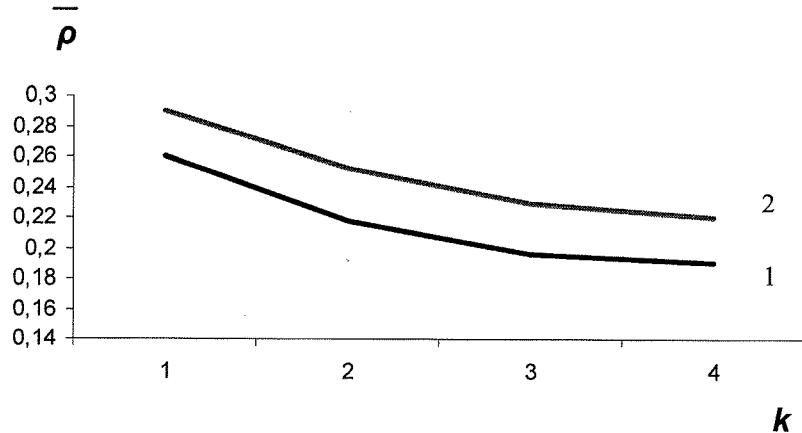


Рисунок 2. – Зависимость оценки вероятности ошибки $\bar{\rho}$ распознавания образов от количества признаков классифицируемых объектов k при $\beta = 0.06$, $m = 7$ и $n = 16$. Кривые 1, 2 соответствуют исследуемому и традиционному непараметрическим алгоритмам распознавания образов.

В четвертой главе рассматривается и исследуется методика восстановления стохастических зависимостей $y = \varphi(x)$, $x \in R^k$ в условиях малых выборок $V = (x^i, y^i, i = \overline{1, n})$ распределённых с неизвестной плотностью $p(x, y)$.

В исследуемой модели используется уравнение вида

$$\bar{y} = \int_{-\infty}^{+\infty} y p\left(\frac{y}{x}\right) dy, \quad (9)$$

являющееся оптимальным оператором в смысле среднеквадратического значения. Здесь $p\left(\frac{y}{x}\right)$ – условная плотность вероятности.

Для того, чтобы обосновано оценивать $y = \varphi(x)$ увеличим объем исходной выборки за счет результатов статистического моделирования.

С этой целью в β -окрестности каждой i -ой ситуации выборки проведем m имитаций случайных величин x_Δ и y_Δ с законами распределения $p(x_\Delta)$ и $p(y_\Delta)$ соответственно и нулевыми математическими ожиданиями. В результате получим статистическую выборку $V_2 = (x^i + x_\Delta^j, y^i + y_\Delta^j, j = \overline{1, m}, i = \overline{1, n})$.

Подставляя в выражении (9) вместо $p(x, y)$ и $p(x)$ оценки плотности вероятности типа 2, получим

$$y(x) = \frac{\sum_{i=1}^n \sum_{j=1}^m (y^i + y_\Delta^j) \Phi\left(\frac{x - x^i - x_\Delta^j}{c}\right)}{\sum_{i=1}^n \sum_{j=1}^m \Phi\left(\frac{x - x^i - x_\Delta^j}{c}\right)}.$$

Многомерный аналог непараметрической регрессии в условиях малых выборок, когда x вектор размерности k , имеет вид

$$y(x) = \frac{\sum_{i=1}^n \sum_{j=1}^m (y^i + y_\Delta^j) \prod_{v=1}^k \Phi\left(\frac{x - x_v^i - x_\Delta^j}{c_v}\right)}{\sum_{i=1}^n \sum_{j=1}^m \prod_{v=1}^k \Phi\left(\frac{x - x_v^i - x_\Delta^j}{c_v}\right)}. \quad (10)$$

Оптимизация статистики (10) по коэффициентам размытости ядерных функций c_j , $j = \overline{1, m}$ и параметрам процедуры размножения выборок производится в режиме «скользящего экзамена» из условия минимума оценки среднеквадратического критерия.

Для достаточно гладких восстанавливаемых функций, плотностей вероятности её значений и аргументов доказаны теоремы об асимптотической несмещённости непараметрической регрессии в условиях малых выборок, что аналитически обосновывает предложенный подход.

По результатам вычислительного эксперимента установлено, что аппроксимационные свойства непараметрических моделей в условиях малых выборок менее чувствительны к помехам в данных и размерности обучающей выборки k по сравнению с непараметрической регрессией. Эмпирическая ошибка непараметрической модели восстановления стохастических зависимостей достоверно отличается от непараметрической регрессии при количестве имитаций $m > 3$.

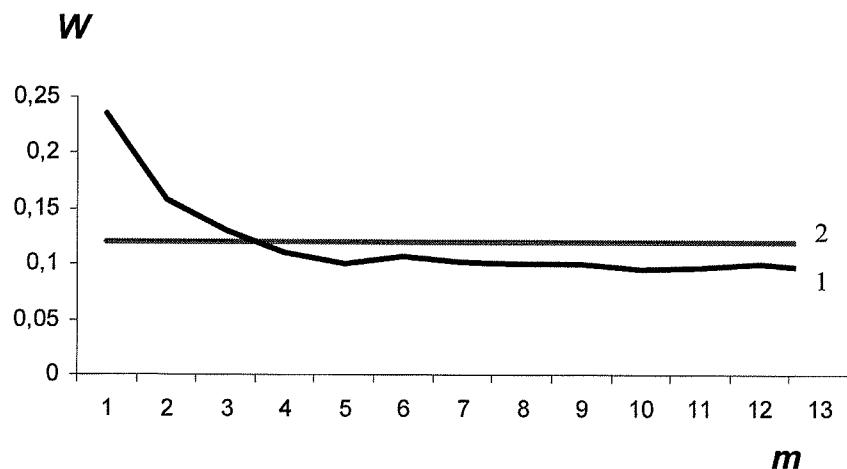


Рисунок 3. – Зависимость средней относительной ошибки аппроксимации функции $y(x) = 1 - x + \exp(-200(x - 0.5)^2)$ от количества имитаций m при $k = 3$, $\beta = 0.1$ и $n = 10$ (кривая 1). Линия 2 - непараметрическая регрессия, восстановленная по исходной выборке объема $n=10$

Достаточное количество имитаций m для достижения максимальной эффективности восстановления составляет 10-11. Увеличение параметра m имитационной модели обеспечивает незначительное снижение ошибки восстановления стохастических зависимостей в условиях малых выборок и увеличивает время, требующееся для работы имитационной модели.

При относительно больших объемах выборок (более 20 элементов) ошибка сходится к ошибке восстановления при помощи традиционной непараметрической регрессии.

В пятой главе приводится описание программного приложения в среде Borland C++ Builder, в котором реализованы методы обработки информации в условиях малых выборок для решения задач восстановления плотности вероятности, восстановления стохастических зависимостей и распознавания образов. Предлагается их применение при исследовании динамики состояния преступности в регионах РФ. Структуру комплекса программ составляют 5 основных блоков (Рисунок 4).

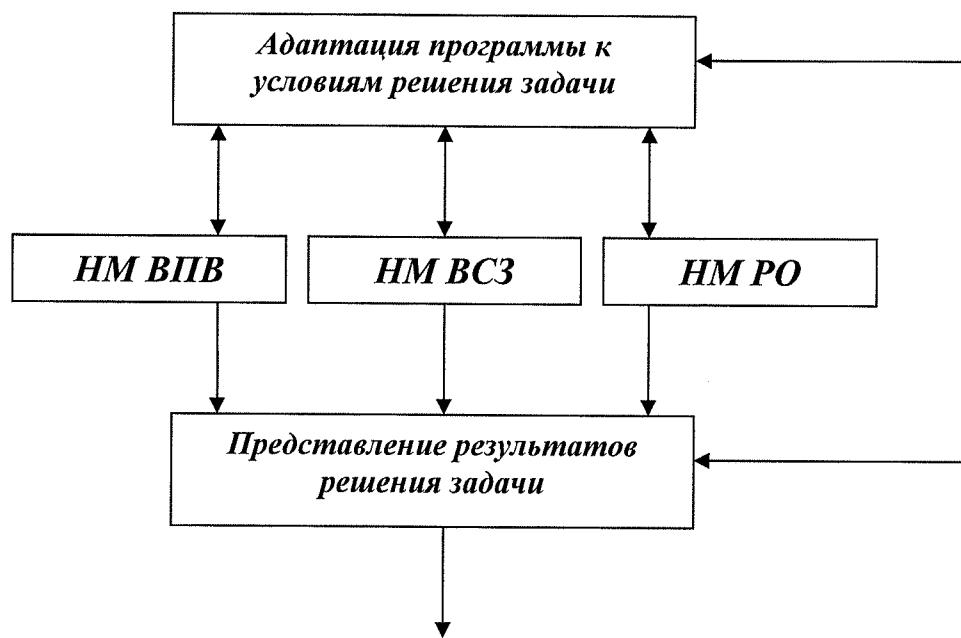


Рисунок 4. - Структура пакета программ «Методы обработки информации в условиях малых выборок» (**НМ ВПВ** - непараметрические методы восстановления плотностей вероятности; **НМ ВСЗ** - непараметрические методы восстановления стохастических зависимостей; **НМ РО** - непараметрические методы распознавания образов).

Под **блоком** в данной работе понимается совокупность взаимосвязанных модулей, предназначенных для решения конкретной функциональной задачи.

Под программным **модулем** понимается некоторая совокупность процедур и функций, связанных между собой определённой функциональной целью. Если модуль имеет самостоятельное значение, т.е. реализует одну из задач, будем называть его блоком.

Блок «Адаптация программы к условиям решения задачи» предназначен для выбора в диалоговом режиме вида задачи, определения условий её решения, указания типов используемых переменных и планирования предварительной обработки данных обучающих выборок.

Его введение придаёт комплексу программ универсальный характер и позволяет использовать его при исследовании объектов различной природы.

Блок «Непараметрические методы восстановления плотности вероятности» предназначен для восстановления неизвестных многомерных плотностей вероятности с использованием разработанных непараметрических методов на основе аппарата имитационного моделирования. Основу блока НМВПВ составляют непараметрические алгоритмы восстановления плотностей вероятности.

Блок «Непараметрические методы распознавания образов» обеспечивает решение задач классификации при малом объеме априорной.

Блок «Непараметрические методы восстановления стохастических зависимостей» предназначен для восстановления неизвестных многомерных стохастических зависимостей с использованием разработанных непараметрических методов восстановления стохастических зависимостей в условиях малых выборок.

Блок «Представление результатов решения задачи» служит для визуализации результатов решения функциональных задач с возможностью сохранения их в виде текстового файла или вывода на печатающее устройство.

Разработанные методы используются в программе «Количественное оценивание состояния преступности в регионах РФ», обеспечивающей обнаружение зависимостей показателей преступности от социально-экономических и демографических условий. Ее применение открывает возможность создания системы поддержки принятия решений при определении эффективности деятельности подразделений правоохранительных органов и выборе профилактических и оперативных мероприятий.

Исходной информацией, необходимой для синтеза моделей, являются данные сборника «Преступность и правонарушения (1998-2002г.г.)»

Программа зарегистрирована в «Отраслевом фонде алгоритмов и программ» (Свидетельство № 6787).

Функциональные возможности комплекса программ:

1. Нормирование исходных статистических данных, которыми являются количество совершенных преступлений определенного вида в регионе за

период на 100 000 жителей, что позволяет ввести понятие состояния преступности с учетом взаимосвязи между относительными показателями видов преступности.

2. Увеличение объема исходных данных при помощи аппарата имитационного моделирования.

3. Объединение территорий региона в группы по данным относительных значений показателей видов преступности в каждый интервал времени их контроля с применением решающего правила 6. Группу (класс) составляют территории, которым соответствуют компактные множества точек в пространстве относительных значений показателей видов преступности.

4. Установление обобщённого коэффициента взаимосвязи между показателями видов преступности для обнаруженных групп территорий. Обобщённый коэффициент взаимосвязи определяется как среднее значение положительных коэффициентов парных взаимосвязей между показателями видов преступности, соответствующих группе территорий региона. Анализ взаимосвязи между различными видами преступности внутри выделенных групп территорий осуществляется с применением моделей типа 10.

5. Анализ развития преступности на территории осуществляется путём исследования временной траектории её состояний. Траектория состояний формируется по данным принадлежности временного ряда показателей видов преступлений на территории. По тенденции изменения состояний преступности появляется возможность оценить эффективность мероприятий, проводимых правоохранительными органами территории.

На основе комплекса программ установлена неоднородность субъектов Российской Федерации в пространстве показателей видов преступлений. Определены 4 группы регионов, достоверно отличающиеся уровнем преступности и обобщенным показателем взаимосвязи между ее видами. Установлено, что данный показатель может служить объективным критерием оценивания состояния преступности в регионе и эффективности борьбы с ней. Его увеличение является благоприятной тенденцией динамики преступности региона и указывает на повышение эффективности работы правоохранительных органов в условиях централизации управления в стране.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

Впервые с позиций принципов имитации систем и методов непараметрической статистики теоретически обоснованы и решены задачи восстановления многомерных плотностей вероятности, стохастических зависимостей и распознавания образов в условиях малых выборок.

Сформулированная цель диссертации достигнута, получены следующие основные результаты и выводы:

1. Развиты теоретические основы построения непараметрических моделей в условиях малых выборок в задачах восстановления многомерных плотностей вероятности, восстановления стохастических зависимостей и распознавания образов, основанные на принципах имитационного моделирования.
2. При увеличении объема исходных данных и уменьшение размера β -окрестности разработанные непараметрические статистики являются асимптотически несмещенными и состоятельными оценками, что обосновывает эффективность предлагаемого подхода синтеза решающих правил в условиях малых выборок.
3. Установлено, что интервал генерирования искусственной обучающей последовательности увеличивается с ростом области определения $p(x)$ и количества имитаций m , снижается по мере увеличения объема n исходной выборки.
4. На основе результатов исследования асимптотических свойств непараметрических оценок плотности вероятности предложена и обоснована методика продолжения случайных последовательностей.
5. Выявлено, что с ростом размерности признаков k исходной выборки достоверное преимущество предлагаемых моделей над традиционными сохраняется;
6. Установлено, что для моделей восстановления стохастических зависимостей и алгоритмов распознавания образов оптимальным условием выбора β -окрестности является условие максимума перекрывания области определения базовой выборки при минимальном перекрывании β -областями друг друга.
7. Определено, что оптимальное значения параметра имитационной процедуры $m \in (10;12)$ при размерности пространства признаков обучающих выборок $k < 4$. Увеличение данного значения приводит к незначительному снижению ошибки моделирования в условиях малых выборок и увеличивает время обработки информации;
8. Созданы информационные средства, реализующие непараметрические методы обработки информации в условиях малых выборок в задаче анализа динамики показателей преступности регионах России и имеющие государственную регистрацию (свидетельство о регистрации № 6787).

Публикации автора по теме диссертации

В изданиях, включенных в список работ, рекомендованных ВАК для публикации результатов диссертаций:

1. Шарков М.А. Синтез и анализ непараметрических методов обработки информации в условиях малых выборок. / М.А. Шарков // Вестник КрасГАУ, 2007. - №1(16)– С. 37-43.

В прочих изданиях:

2. Шарков М.А. Количественное оценивание состояния преступности в регионах [государственная регистрация №ОФАП:6787] / М.А. Шарков // Отраслевой фонд алгоритмов и программ. Москва, 2006.
3. Шарков М.А. Непараметрическая оценка плотности вероятности в условиях малых выборок. / М.А. Шарков // Молодежь и современные информационные технологии: Сборник трудов IV научно-практической конференции студентов, аспирантов и молодых ученых. - Томск: ТПУ, 2006.–С. 166-168.
4. Шарков М.А. Непараметрические методы распознавания образов и восстановления стохастических зависимостей в условиях малых выборок / М.А. Шарков // Вестник НИИ СУВПТ, 2006. - №7(21).–С. 170-176.
5. Шарков М.А. Статистические модели анализа динамики состояния преступности / М.А. Шарков, А.А. Лапко // Молодежь и современные информационные технологии: Сборник трудов V научно-практической конференции студентов, аспирантов и молодых ученых. - Томск: ТПУ, 2007.–С.195-197.
6. Шарков М.А. Синтез и анализ непараметрических алгоритмов восстановления стохастических зависимостей в условиях малых выборок / М.А. Шарков // Электронный журнал «Исследовано в России», 2007.–С.742-753. <http://zhurnal.ape.relarn.ru/articles/2007/071.pdf>.

Подписано к печати «6» 09 2007 г.

Тираж 100 экз. Заказ № 734

Отпечатано в типографии ПИ СФУ

660036, Красноярск, ул. Киренского, 26