На правах рукописи

Ступина Татьяна Александровна

ПОСТРОЕНИЕ ЛОГИКО-ВЕРОЯТНОСТНОЙ МОДЕЛИ ПРОГНОЗИРОВАНИЯ СИСТЕМЫ РАЗНОТИПНЫХ ПЕРЕМЕННЫХ

05.13.18 – математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Новосибирск-2006

Работа выполнена в Институте математики им. С.Л. Соболева
Сибирского Отделения Российской Академии Наук

Научный руководитель:	доктор технических наук, профессор Г.С. Лбов
Научный консультант:	доктор физико-математических наук, Ю.А. Зуев
Официальные оппоненты:	доктор технических наук, доцент А.С. Родионов
	кандидат физико-математических наук, доцент И.А. Пестунов
Ведущая организация:	Новосибирский государственный технический университет

Защита состоится « 14 » ноября 2006 г. в 15-00
На заседании диссертационного совета Д 003.061.02 при Институте вычислительной математики и математической геофизики СО РАН по адресу: 630090, Новосибирск, пр. ак. Лаврентьева,6.

С диссертацией можно ознакомиться в библиотеке Института вычислительной математики и математической геофизики СО РАН.

Автореферат разослан « 9 » октября 2006 г.

Ученый секретарь
диссертационного совета
доктор физико-математических наук  С.Б. Сорокин

Общая характеристика работы

Актуальность темы. Одним из важных направлений в области информатики является решение задач построения решающих функций распознавания и прогнозирования на основе анализа эмпирической информации, заданной в виде таблиц данных, временных рядов и экспертных знаний. Методы построения решающих функций с успехом применяются в различных научных исследованиях при решении задач в таких областях, как экология, медицина, социология, археология и т.д. К настоящему времени разработано большое количество методов построения решающих функций, основанных на различных идеях, гипотезах и принципах. Однако существующие подходы и методы построения решающих функций в задачах анализа многомерной эмпирической информации ориентированы, в основном, на случай одной целевой переменной (например, задача распознавания образов, регрессионного анализа). Случай одновременного прогнозирования нескольких переменных рассматривался, например, для количественных переменных в задачах многооткликовой регрессии. Поэтому работы в данной области остаются актуальными.

При решении задач анализа данных с использованием эмпирического материала ограниченного объема важными в теоретическом и практическом плане являются проблема определения качества метода и исследование его зависимости от сложности распределения, сложности используемого класса решающих функций и объема обучающей выборки. В работах данного направления понятие сложности распределения, сложности класса решающих функций формализуется по-разному. Результаты исследований дают возможность строить наилучшую решающую функцию при ограниченном объеме обучающей выборки с учетом сложности распределения, сложности класса решающих функций. Основные результаты решения этой проблемы получены в области построения решающих функций распознавания (Ш.Ю. Раудис, Г.С. Лбов, Н.Г. Старцева, В.Б. Бериков и др.).

Необходимость разработки методов прогнозирования системы разнотипных переменных и исследование их качества обуславливается существованием достаточно широкого круга прикладных задач в естественнонаучных областях. В качестве примера можно привести задачу выявления взаимосвязи между характеристиками экологической обстановки и характеристиками здоровья населения региона. В этой задаче необходимо по характеристикам экологической ситуации предсказать разнотипный набор характеристик здоровья населения. Как показывают теоретические и экспериментальные исследования (Г.С. Лбов, Н.Г. Старцева), наиболее подходящим классом функций для анализа разнотипной информации является класс логических решающих функций, который послужил основой при построении логико-вероятностной модели и стал средством исследований, проделанных в работе.

Цель работы заключается в разработке метода построения логико-вероятностной модели прогнозирования системы разнотипных переменных и способа его исследования, в частности, исследование методов построения кусочно-линейных регрессионных функций.

Методы исследований. В работе используется аппарат теории вероятностей, математической статистики, теории статистических решений, линейного регрессионного анализа, распознавания образов.

Научная новизна. В работе впервые получены следующие результаты:

- разработан способ оценивания качества метода прогнозирования системы разнотипных переменных (ПСРП);

- предложен метод построения логико-вероятностной модели прогнозирования системы разнотипных переменных;
- получены зависимости, позволяющие определить влияние типа переменной (с упорядоченным и неупорядоченным набором значений) на качество решения при ПСРП в условиях малой выборки;
- получены зависимости качества метода ПСРП от сложности распределения, сложности класса решающих функций и объема выборки;
- для порогового метода построения кусочно-постоянных решающих функций при заданном классе распределений получена нижняя оценка его качества в зависимости от сложности класса решающих функций и объема выборки;
- предложен критерий обнаружения значимого подмножества переменных МНК-метода построения линейной регрессионной функции.

Практическая ценность результатов работы. Теоретические исследования и методы, предложенные в данной работе, позволяют решать прикладные задачи выбора значимого подмножества переменных в линейном регрессионном анализе, задачи прогнозирования системы разнотипных переменных, что существенно расширяет круг прикладных задач анализа данных, анализа многомерных временных рядов. Результаты были использованы при решении прикладных задач из области медицины и гидрологии. Программная реализация разработанных методов является эффективным инструментом в статистической обработке данных и может быть применена в научно-исследовательских работах в области медицины, экологии, гидрологии и других естественнонаучных областях.

На защиту выносятся:

Разработка способа оценивания качества метода прогнозирования системы разнотипных переменных.

Метод построения логико-вероятностной модели прогнозирования системы разнотипных переменных, основанный на предложенном критерии, с учетом влияния разнотипности пространства.

Результаты анализа зависимости качества метода ПСРП от сложности распределения, сложности решающих функций, объема выборки.

Результаты оценивания качества порогового метода построения кусочно-постоянных функций в зависимости от сложности класса решающих функций и объема выборки при известном классе распределений.

Критерий обнаружения значимого подмножества переменных МНК-метода построения линейной многомерной регрессионной функции.

Апробация работы. Основные положения работы докладывались и обсуждались на Конгрессе по индустриальной и прикладной математике (ИНПРИМ-98, Новосибирск); Всероссийских конференциях «Математические методы распознавания образов» (ММРО-99, 2001, 2003, 2005, Москва); VI Международной конференции «Современные методы математического моделирования природных и антропогенных катастроф» (2001, Красноярск); Международной конференции «Искусственный интеллект» (2002, 2004, Алушта); Международной конференции «Информационные системы и технологии » (IST'2002, 2004, Минск); VII и VIII Международной научной конференции (PRIP-2003, 2005, Минск); Всероссийской конференции «Математические и информационные технологии в энергетике, экономике, экологии» (2003, Иркутск); научной немецко-российской школе-семинаре «Распознавание образов и изображений» (2003, Алтай); Международной конференции «Knowledge-Dialogue-Solution» (KDS'2005, 2006 Bulgaria).

Связь с государственными программами. Работа выполнена в рамках проектов № 95-01-00930а, 98-01-00673, 01-01-00839, 04-01-00858, поддержанных РФФИ; Интеграционного проекта СО РАН №13.10 «Анализ и моделирование экстремальных гидрологических явлений».

Публикации. По теме диссертации автором опубликована 21 работа.

Структура и объем работы. Диссертация объемом 155 страниц состоит из введения, четырех глав, заключения, списка литературы из 92 наименований.

Содержание работы

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель работы и задачи исследований, приведены основные положения, выносимые на защиту, дано краткое изложение по главам.

Первая глава является вводной и содержит краткий обзор существующих методов построения решающих функций (моделей) и способов определения их качества в задачах распознавания образов и регрессионного анализа. Рассматривается общая постановка задачи восстановления зависимости, для которой задачи распознавания образов и регрессионного анализа являются частными случаями.

Пусть объект a из генеральной совокупности Γ описывается набором (системой) переменных $(X_1,...,X_n;Y_1,...,Y_m)$, которому соответствует набор значений $(x,y)=(x_1,...,x_n;y_1,...,y_m)$ в многомерной разнотипной области $D=D_X\times D_Y$, где $D_X=\prod_{j=1}^{n}D_{X_j}$, D_{X_j} - множество допустимых значений переменной X_j, $D_Y=\prod_{j=1}^{m}D_{Y_j}$, D_{Y_j} - множество допустимых значений переменной Y_j. Оба набора переменных могут быть произвольных типов (количественные, порядковые, номинальные). Пусть задано вероятностное пространство $<D,\mathrm{B},P>$, где $P=P[D]$ - вероятностная мера на борелевской σ-алгебре B (такую меру будем обозначать через *c* и называть стратегией природы). Под решающей функцией *f* понимается соответствие между набором значений переменных $(x_1,x_2,...,x_n)$ и набором значений прогнозируемых (целевых) переменных $(y_1,y_2,...,y_m)$, т.е. $f:D_X\to\wp(D_Y)$, где $\wp(D_Y)$ - область значений функции. Качество решающей функции оценивается с помощью функционала качества $F(c,f)$. Если $F(c,f^*)=\inf_{f\in\Phi}F(c,f)$ (либо $F(c,f^*)=\sup_{f\in\Phi}F(c,f)$), то f^*- оптимальная решающая функция в заданном классе. Обозначим через Φ° класс всех измеримых функций, тогда $\Phi\subseteq\Phi^\circ$.

При *m*=1, $\wp(D_Y)=D_Y$ имеем задачу распознавания образов для номинального типа переменной ($D_Y=\{\omega_1,...,\omega_k\}$, ω_i - имя образа) и задачу восстановления зависимости (регрессионной функции) для непрерывного типа переменной ($D_Y=R$). Критерий качества решающей функции в этих случаях, как правило, определяется через функцию риска $F(c,f)=R(c,f)=\int_D L(f,y)dP[D]$ (ожидаемые потери для решающей функции *f*). Функция потерь $L(f,y)$ задается в зависимости от типа переменной (например, $L(f,y)=\begin{cases}0, & \text{при } f=y\\ 1, & \text{при } f\neq y\end{cases}$ в распознавании образов, $L(f,y)=(y-f)^2$ в регрес-

сионном анализе) и от специфики прикладной задачи (например, матрицей потерь). Оптимальной решающей функцией распознавания в классе Φ° всех измеримых является байесовская решающая функция f_B, для которой $P_{f_B} = \inf_{f\in\Phi^\circ} R(c,f)$, в регрессионном анализе – функция регрессии $f(x) = \int_{D_Y} yp(y/x)dy$. В главе 3 для многомерного случая, $m \geq 1$, рассматривается решающая функция прогнозирования области и вводится соответствующий критерий качества.

Заметим, что для оценивания качества решающей функции кроме риска может использоваться такое понятие как трудоемкость алгоритма вычисления решающей функции (время принятия решения для фиксированного объекта). Однако риск является наиболее важным, поэтому при определении качества именно он и рассматривается.

При анализе эмпирической информации, представленной выборкой v_N ограниченного объема *N*, методом Q строится выборочная решающая функция $\bar{f} = Q(v_N)$ из класса Φ. Под методом будем понимать отображение $Q : \{v_N\} \to \Phi$ и сам способ его построения (алгоритм). Необходимо определить качество метода и исследовать его в зависимости от сложности стратегии природы, сложности класса решающих функций и объема обучающей выборки. Результаты исследований позволят судить о возможности применения (области применимости) метода (получения хороших решений) при анализе данных ограниченного объема.

На практике, как правило, стратегия природы неизвестна, поэтому принимаются предположения о виде распределения (ограничения на класс распределений Λ) либо о постулируемой модели (ограничения на класс решающих функций Φ), либо о том и другом. Разнообразие сделанных предположений указывает на существование достаточно большого количества исследований, проводимых при изучении качества метода.

При заданной стратегии природы определим качество метода через ожидаемый по выборкам функционал качества $\mathbf{E}_{V_N} F(c,\bar{f})$. Также может быть вычислена степень неадекватности класса решающих функций к стратегии природы $\gamma(c) = \inf_{f\in\Phi} F(c,f) - \inf_{f\in\Phi^\circ} F(c,f)$, $\Phi \subseteq \Phi^\circ$ и степень отклонения от оптимального в классе $\kappa(c) = \mathbf{E}_{V_N} F(c,\bar{f}) - \inf_{f\in\Phi} F(c,f)$ для метода *Q*. Величину $\kappa(c)$ можно рассматривать как некоторую дополнительную меру качества метода. Исследование метода обучения сводится к нахождению функциональной зависимости $\mathbf{E}_{V_N} F(c,\bar{f})$ от сложности M_Φ класса решающих функций, в котором работает метод, и от объема обучающей выборки, т.е. $g_1(c, M_\Phi, N)$. Такой подход был применен Ш. Ю. Раудисом, Г. С. Лбовым и др. к задаче распознавания образов, когда функционал качества определялся вероятностью ошибки.

Если на множестве Λ всех стратегий природы задано распределение $P[\Lambda]$, то качество метода определяется как усредненный по стратегиям природы и выборкам функционал качества $\mathbf{E}_c\mathbf{E}_{V_N} F(c,\bar{f})$. Исследование качества метода обучения в данном варианте сводится к нахождению функциональной зависимости величины

$\mathbf{E}_c\mathbf{E}_{V_N}F(c,\bar{f})$ от сложности M_Λ класса стратегий природы, сложности M_Φ класса решающих функций и объема обучающей выборки, т. е. $g_2(M_\Lambda, M_\Phi, N)$.

Если стратегия природы неизвестна, то о качестве метода построения решающих функций, вообще говоря, судить сложно, поскольку всегда найдется стратегия, при которой данным методом может быть получена плохая решающая функция.

Отдельный вопрос, который затрагивается в третьей главе, – оценивание качества решающей функции, построенной по фиксированной выборке при неизвестной стратегии природы. Для его решения в литературе существует хорошо известный подход Вапника-Червоненкиса, основанный на определении доверительной границы ε отклонения риска от эмпирического риска $\overline{F}(\bar{f})$. Для получения аналитических оценок данным способом необходимо знание или возможность вычисления ёмкостной характеристики класса решающих функций, в котором строятся решения. Однако многие используемые на практике методы обладают бесконечной ёмкостью либо трудно вычислимой. В диссертационной работе для некоторых параметрических семейств стратегий природы эмпирически было оценено смещение эмпирического функционала качества $\varepsilon_N(c) = \mathbf{E}_{V_N}F(c,\bar{f}) - \mathbf{E}_{V_N}\overline{F}(\bar{f})$, которое позволяет судить о качестве решающей функции по значению эмпирического функционала.

Одновременно с определением качества метода возникают вопросы о том, как вводить ограничения на класс распределений, на класс решающих функций, как определять M_Λ сложность класса стратегий и M_Φ сложность класса решающих функций, каков должен быть достаточный объем обучающего материала *N* для достижения заданного качества. Многие из этих вопросов остаются открытыми до сих пор.

В работе автором рассматриваются методы построения решающих функций из заданного класса, основанные на минимизации (максимизации) эмпирического функционала качества. Для построения решающих функций в разнотипном пространстве был использован класс логических решающих функции (ЛРФ), описание которого приводится в параграфе 5.

Во второй главе проведено исследование качества метода построения кусочно-постоянных регрессионных функции при заданном классе распределений в зависимости от сложности решающей функции и объема выборки. Предложен критерий обнаружения значимого подмножества (набора) переменных МНК-метода построения многомерной линейной регрессионной функции и проведено исследование его качества.

Как частный случай задачи, сформулированной в первой главе, рассматривается одномерная кусочно-постоянная регрессионная модель, *n*=1, *m*=1, $D_X = [0,1]$, $\wp(D_Y) = D_Y = \mathbf{R}$. Сложность решающей функции определяется числом M' областей разбиения, сложность стратегии природы при фиксированном равномерном распределении внутри каждой подобласти – числом *M* подобластей постоянства.

По выборке $v_N = \{x^i, y^i\}_{i=1,N}$ пороговым методом $Q(v_N)$ строится кусочно-постоянная решающая функция $\bar{y} = \bar{f}(x) = \sum_{k=1}^{M'} \hat{s}_k \hat{\psi}_k(x)$, где $x \in D_X$, $y \in D_Y$, $\hat{\psi}_k(x) = \begin{cases} 1, \text{при } x \in [\hat{b}_{k-1}, \hat{b}_k) \\ 0, \text{при } x \notin [\hat{b}_{k-1}, \hat{b}_k) \end{cases}$, $k = 1,...,M'$, т.е. $\bar{f} \in \Phi_{M'}$, $M_\Phi = M'$. Пороговый ме-

тод осуществляет расстановку выборочных границ $\hat{b}_k$ в случае преодоления элементами выборки некоторого порога $h_\circ$, т.е. $\max\limits_{x_i, x_j \in [\hat{b}_{k-1}, \hat{b}_k)} | y_i - y_j | \le 2h_\circ$. При заданной стратегии природы $c = \{ p(x, y), p(x) \in U[0,1], p(y / x \in [b_{t-1}, b_t)) \in U[s_t - h, s_t + h], h \in D_Y, s_t \in D_Y, [b_{t-1}, b_t) \in \alpha, \alpha \in \Psi_M, t = 1,..., M \}$ сложности $M_\Lambda = M$, где Ψ_M - множество разбиений области D_X на M подобластей, $U[\gamma_1, \gamma_2]$ - класс равномерных на отрезке $[\gamma_1, \gamma_2]$ распределений, функционал качества выборочной решающей функции определяется через риск и равен $F(c, \bar{f}) = \sum_{t=1}^{M} (b_t - b_{t-1})(s_t - \bar{f}(x))^2 + \frac{h^2}{3}$. Наилучшим является решение, при котором данный критерий принимает минимальное значение.

Будем рассматривать метод $Q(v_N)$ расстановки $M' = M$ эмпирических границ $\hat{b}_k$ такой, что решение принимается в виде $\bar{f}(x) = \hat{s}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} y^i$ (среднее значение в интервале), где N_k - число точек выборки попавших в интервал $[\hat{b}_{k-1}, \hat{b}_k)$. С применением аппарата порядковых статистик получена: 1) нижняя оценка качества порогового метода $\mathrm{E}_{V_N} F(c, \bar{f}) \ge \sum_{k=1}^{M-1} (s_k - s_{k+1})^2 \, | b_k - (\frac{i_k}{N+1} + \frac{1}{2(N+1)}) | + \frac{h^2}{3(N+1)} (M + \frac{1}{2N_M} + \frac{1}{2N_1}) + \frac{h^2}{3}$, где i_k - номер порядковой статистики в ранжированном ряде $x_{i_1} \le x_{i_2} \le ... \le x_{i_k} \le ... \le x_{i_N}$; 2) степень отклонения от оптимального решения в классе $\kappa(c, N) = \mathbf{E}_{V_N} F(c, \bar{f}) - \frac{h^2}{3}$; 3) для стратегий природы из класса $\Lambda'(M) = \{ c \mid E^k = [b_k, b_{k+1}), b_k = \frac{k}{M}, s^* = \max\limits_{k=0,...,M-1} (s_k - s_{k+1})^2, \frac{\sqrt{s^*}}{2} > h_\circ \}$ верхняя оценка качества метода $\mathbf{E}_{V_N} F(c, \bar{f}) \le \frac{h^2}{3} (1 + \frac{M}{N})$.

В качестве дополнительного результата найдена плотность распределения границ областей разбиения (*утверждение 2.6*).

В третьем параграфе рассматривается еще один частный случай задачи в общей постановке, когда $m = 1$, $\wp(D_Y) = D_Y$, $D_X = D_Y = \mathbf{R}$, $M_\Phi = n, M_\Lambda = (k, \sigma^2)$, т.е. задача обнаружения значимого подмножества из k переменных в классической линейной многомерной регрессионной модели. Под сложностью решающей функции в данном случае рассматривается число наблюдаемых переменных (регрессоров). Модель в матричном виде: $Y = \theta + \varepsilon = W\beta + \varepsilon = (X, Z)(\beta_I, \beta_{II}) + \varepsilon = \beta_\circ + \beta_1 X_1 + ... + \beta_k X_k + \beta_{k+1} Z_{k+1} + ... + \beta_n Z_n + \varepsilon$, где Y - прогнозируемая переменная, X_i - значимая переменная (i=1,...,k), Z_j - незначимая переменная, т.е. $\beta_j = 0$ (j=k+1,...,n), ε - шум ($p(\varepsilon) \in N(0, I\sigma^2)$). Значимый и незначимый наборы переменных становятся практически неразличимы при принятии решения по выборкам малого объема (например, $\frac{N}{n+1} < 10$) в случае сильной зашумленности (уровень шума определяется его дисперсией). Для этой модели функция риска в точке $w = (1, x_1, ..., x_k, z_{k+1}, ..., z_n)$ определяется как среднеквадратичная ошибка прогноза, т.е.

$F(c,\bar{f}) = \mathbf{E}L(\hat{Y},Y) = \mathbf{E}[(\hat{Y} - \mathbf{E}[Y])^2]$, где $\bar{f}(w) = \hat{Y} = \hat{\beta}_\circ + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n z_n$, $\hat{\beta}_i$ - МНК-оценка параметра β_i. Показано, что $F(c,\bar{f}) = \mathbf{D}[\hat{Y}_p] + (\varsigma_p - \theta)^2$, где $\varsigma_p = E[\hat{Y}_p]$, p-количество параметров, оцениваемых на подмодели (линейная регрессионная функция от p переменных, $p<n$). На полной модели (от n переменных) индекс p опущен. Второе слагаемое указывает на степень неадекватности выбранной модели по отношению к истинной θ. Величина ошибки прогноза при исключении из модели значимого набора по отношению к величине ошибки на полном наборе переменных изменяется значительно больше, чем при исключении незначимого набора. Эта идея используется в критерии. Для известной дисперсии шума σ^2 критерий представляется отношением

$$\mu_w = \frac{\Delta D_н}{\Delta D_з} = \frac{EL(\hat{Y}_{n-k},Y) - EL(\hat{Y},Y)}{EL(\hat{Y},Y) - EL(\hat{Y}_k,Y)} = \frac{((z'(Z'Z)^{-1}Z'X - x')\beta)^2 - \sigma^2(w'(W'W)^{-1}w - z'(Z'Z)^{-1}z)}{\sigma^2(w'(W'W)^{-1}w - x'(X'X)^{-1}x)}.$$

Если $\mu_w \le 1$, то набор из k переменных, предполагаемый значимым, на самом деле не является полным набором значимых переменных (либо обрабатываемые данные обладают недостаточной информацией). Если $\mu_w > 1$, то набор из k переменных, предполагаемый значимым, действительно таковым является. В случае неизвестного распределения для построения выборочной оценки введенного критерия предлагается использовать статистику Маллоуса: $\hat{\mu}_w = \frac{\Delta \hat{D}_н}{\Delta \hat{D}_з} = \frac{C_{n-k+1} - C_{n+1}}{C_{n+1} - C_{k+1}} = \frac{RSS_{n-k+1} - u_1 RSS_{n+1}}{u_2 RSS_{n+1} - RSS_{k+1}}$, где C_i -статистика Маллоуса на модели от i параметров, $u_1 = 1 + \frac{2k}{N-n-1}$, $u_2 = 1 + \frac{2(n-k)}{N-n-1}$, RSS_p - остаточная сумма квадратов на модели из p переменных.

На модельных примерах показана эффективность предложенного метода обнаружения истинного подмножества значимых переменных. С этой целью было проведено моделирование ста выборок фиксированного объема *N=10, 20, 50, 70, 100* при заданном уровне шума σ и заданном числе k значимых переменных, определяющих стратегию природы.

В третьей главе приводится постановка задачи прогнозирования системы разнотипных переменных (ПСРП), разработан способ оценивания качества метода ПСРП, который включает задание класса стратегий природы и функционала качества решающих функций. Предложен метод ПСРП в классе логических решающих функций, основанный на предложенном эмпирическом критерии, и представлены результаты исследования зависимости его качества от сложности стратегии природы, сложности класса решающих функций и объема выборки. Прогнозирование осуществляется в классе функций, значения которых представимы областями в многомерном разнотипном пространстве переменных. Отмечается, что задачи распознавания образов и регрессионного анализа являются частными случаями предложенной постановки.

Пусть определено вероятностное пространство $< D, \mathrm{B}, P >$, где $D = D_X \times D_Y$, $\dim D_X = n$, $\dim D_Y = m$, B – борелевская σ-алгебра на D, $P[D]$- вероятностная мера на B (будем обозначать через c). На разнотипном (номинально-вещественном) пространстве $D = D_н \times D_в$ определим меру μ так, что для любого $E \in \mathrm{B}$, $E = \bigcup_{j=1}^{|E_н|} E_в^j \times \{z^j\}$, $\mu(E) = \sum_{j=1}^{|E_н|} \frac{\mu(E_в^j)}{|D_н|\mu(D_в)}$, где $E_н$ - проекция множества E на простран-

ство номинальных переменных $D_н$, z^j - элемент $E_н$, $E_в^j$ - соответствующая элементу z^j область в $D_в$, $\mu(E_в^j)$ - лебегова мера множества $E_в^j$. Для любого подмножества подпространств D_X либо D_Y мера μ задается аналогичным образом. Пусть $\Phi^\circ = \{f : D_X \to \wp(D_Y),\ \wp(D_Y) = 2^{D_Y}\}$ - класс функций с областью определения D_X и со значениями, представленными в виде произвольных множеств $E_y \subseteq D_Y$ (будем обозначать $f(x) = E_y$). Φ° такой, что существует функционал $F(c,f)$, $F(c,f) = \int_{D_X}(P(y \in E_y / x) - \mu(E_y))dP(x)$, где $P(y \in E_y / x)$ - вероятность события $\{y \in E_y\}$ при фиксированном x (в дальнейшем обозначается $P(E_y / x)$), $\wp(D_Y)$ - множество всех подмножеств области D_Y.

Задача прогнозирования системы разнотипных переменных (ПСРП) состоит в том, чтобы для произвольного объекта a из генеральной совокупности Γ по известным значениям переменных $X_1, X_2, ..., X_n$ из области D_X предсказать некоторое множество E_y значений системы целевых (прогнозируемых) переменных $Y_1, Y_2, ..., Y_m$ из области D_Y. Для предсказания необходимо построить такую решающую функцию f из заданного класса $\Phi \subseteq \Phi^\circ$, что $f^* = \arg\max_{f \in \Phi} F(c,f)$.

Утверждение 3.1. Для произвольной стратегии природы c функционал качества $F(c,f)$ представим через функцию риска как $1 - R(c,f) = \int\limits_{D_X}\int\limits_{D_Y}(1 - L(y, f(x)))p(x,y)dxdy$ с потерями вида $L(y,f) = \begin{cases} \mu(E_y), & y \in E_y \\ 1 + \mu(E_y), & y \notin E_y \end{cases}$.

Ряд свойств данного критерия формулируется в виде утверждений и следствий.

Утверждение 3.2. При распознавании k образов решающей функцией f $F(c,f) = \frac{k-1}{k} - P_f$, P_f - вероятность ошибки распознавания.

Утверждение 3.3. В регрессионном анализе оптимальная решающая функция $f_\circ = \arg\max\limits_{f \in \Phi} F(c,f)$, $f_\circ = E_y = [\mathbf{E}(y/x) - \delta_1, \mathbf{E}(y/x) + \delta_2]$, $\mathbf{E}(y/x)$ - оптимальная регрессионная функция, доставляющая минимум функции риска, $\delta_1 \in D_Y$, $\delta_2 \in D_Y$.

Предлагается рассматривать решение задачи ПСРП в классе логических решающих функций Φ_M. Для задачи ПСРП класс ЛРФ определяется следующим образом: $\Phi_M = \{f \in \Phi_\circ \mid f \sim < \alpha, r(a) >, \alpha \in \Psi_M, r(\alpha) \in R_M\}$ (знак '~' обозначает соответствие паре $< \alpha, r(a) >$ символьного знака функции *f*), где Ψ_M -множество всевозможных разбиений $\alpha = \{E_X^1, ..., E_X^M \mid E_X^t = \prod_{j=1}^n E_{X_j}^t,\ E_{X_j}^t \subseteq D_{X_j}, t = \overline{1,M}, \bigcup E_X^t = D_X\}$ области D_X на *M* непересекающихся областей, R_M - множество всевозможных решений $r(\alpha) = \{E_y^1, ..., E_y^M \mid E_y^t \in \Im_{D_Y}, t = \overline{1,M}\}$, $\Im_{D_Y}$ - множество всевозможных *m*-мерных интервалов. Сложность класса ЛРФ определяется параметром *M* в случае одновариантного предсказания (решение представляется в форме: если $x \in E_X^t$, то

$y \in E_y^t$), $M_\Phi = M$, и набором $(k_1,...,k_M)$ в случае многовариантного предсказания, когда $E_y^t = \bigcup_{i=1}^{k_t} E_y^i$, $t = 1,...,M$ и $E_y^i \cap E_y^j = \varnothing$ для $i \neq j$ (решение представляется в форме: если $x \in E_X^t$, то $y \in E_y^1 \vee E_y^2 \vee ... \vee E_y^{k_t}$). В работе рассматривается случай $M_\Phi = M$.

Утверждение 3.4. Если $f \in \Phi_M$, то $F(c,f) = \sum_{t=1}^{M} p_x^t (p_{y/x}^t - \mu_y^t)$, где $p_x^t = P(E_X^t)$, $p_{y/x}^t = P(E_y^t / E_X^t)$, $\mu_y^t = \mu(E_y^t)$, $E_X^t \in \alpha$, $E_y^t \in r(\alpha)$.

Следующее утверждение показывает свойство универсальности класса ЛРФ и возможность его применения без ограничения общности на вид решающих функций.

Утверждение 3.5. Для любой функции $f \in \Phi^\circ$ и $\varepsilon > 0$ существует M и некоторая ЛРФ $f_M \in \Phi_M$ такая, что $|F(c,f) - F(c,f_M)| \leq \varepsilon$.

Определение 3.1. Будем говорить, что стратегия природы c принадлежит классу $L_\varepsilon(M)$, если существует $f \in \Phi_M$ такая, что $|F(c,f) - F(c,f_\circ)| \leq \varepsilon$ для некоторого малого ε. Стратегия природы *c* имеет сложность *M*.

Определение 3.2. Определим стратегию природы c_M следующим образом: $c_M = \{p^t(x,y) = p_x^t p_{y/x}^t = P(x \in E_X^t) P(y \in E_Y^t / x \in E_X^t), t = 1,...,M\}$ и

1) $\sum_{t=1}^{M} p_x^t = 1$; 2) $P(E_Y^t / E_X^t) = p_{y/x}^t$, 3) $P(\overline{E}_Y^t / E_X^t) = 1 - p_{y/x}^t$, где $E_X^t \in \alpha$, $E_Y^t \in r(\alpha)$, $<\alpha, r(\alpha)> \in \Phi_M$, 4) $\forall\ A_X \subseteq E_X^t$ $P(A_X) = p_x^t \frac{\mu(A_X)}{\mu(E_X^t)}$ и $\forall\ A_Y \subseteq E_Y^t$

$P(A_Y / E_X^t) = p_{y/x}^t \frac{\mu(A_X)}{\mu(E_Y^t)}$. Будем говорить, что данная стратегия порождается функцией из класса Φ_M. *Замечание*: Стратегия c_M, порожденная функцией из класса Φ_M, принадлежит классу $L_\varepsilon(M)$.

Утверждение 3.6. Для произвольной $\tilde{f} \in \Phi_{M'}$ сложности M' при фиксированной стратегии $c_M \in L_\varepsilon(M)$ сложности M выполняется $F(c_M, \tilde{f}) = F(\tilde{\alpha}) = \sum_{t'=1}^{M'} \tilde{p}_x^{t'} \rho^{t'} = \sum_{t'=1}^{M'} \tilde{p}_x^{t'} (\tilde{p}_{y/x}^{t'} - \mu_y^{t'})$, где $\tilde{f} \sim <\tilde{\alpha}, r(\tilde{\alpha})>$,

$\tilde{\alpha} = \{\tilde{E}_X^1, ..., \tilde{E}_X^{t'}, ..., \tilde{E}_X^{M'}\}$, c_M порождается $f \sim <\alpha, r(\alpha)>$, $\alpha = \{E_X^1, ..., E_X^t, ..., E_X^M\}$,

$\tilde{p}_x^{t'} = P(x \in \tilde{E}_X^{t'}) = \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{E}_X^{t'} \cap E_X^t)}{\mu(E_X^t)}$,

$\tilde{p}_{y/x}^{t'} = \frac{1}{\tilde{p}_x^{t'}} \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{E}_X^{t'} \cap E_X^t)}{\mu(E_X^t)} \left(p_{y/x}^t \frac{\mu(\tilde{E}_Y^{t'} \cap E_Y^t)}{\mu(E_Y^t)} + (1 - p_{y/x}^t) \frac{\mu(\tilde{E}_Y^{t'}) - \mu(\tilde{E}_Y^{t'} \cap E_Y^t)}{1 - \mu(E_Y^t)} \right)$.

Замечание. Если стратегия c_M такова, что для некоторого t множество E_Y^t совпадает со всей областью D_Y, то $\tilde{p}_{y/x}^{t'} = \frac{1}{\tilde{p}_x^{t'}} \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{E}_X^{t'} \cap E_X^t)}{\mu(E_X^t)} p_{y/x}^t \frac{\mu(\tilde{E}_Y^{t'} \cap E_Y^t)}{\mu(E_Y^t)}$.

Следствие 3.6.1. Для $\tilde{f} \in \Phi_{M'}$ $P(\overline{\tilde{E}}_Y^{t'} / \tilde{E}_X^{t'}) = 1 - \tilde{p}_{y/x}^{t'}$.

Следствие 3.6.2. Если $M=1$ и $E_Y^1 = D_Y$, то $F(c_1, f) = 0$.

Следствие 3.6.3. Если $\tilde{f} \in \Phi_{M'}$ и стратегия c_1 порождается f, для которой $E_Y^1 = D_Y$, то $F(c_1, \tilde{f}) = 0$.

Следствие 3.6.4. Если $\tilde{f} \in \Phi_1$ и $\tilde{E}_Y^1 = D_Y$, то $F(c_M, \tilde{f}) = 0$ для $M \geq 1$.

Утверждение 3.7. Для произвольной стратегии $c_M \in L_\varepsilon(M)$ (порожденной $f_\circ \in \Phi_M$) сложности M и решающей функции $f \in \Phi_M$ точность приближения функционала качества оценивается по формуле: $|F(c, f_\circ) - F(c, f)| \leq |F(c, f_\circ)| \Delta_M$, где $\Delta_M = \int_{D_X} (\mu(E^1 \overline{E}^2) \frac{1}{\mu(E^1)} + \mu(\overline{E}^1 E^2) \frac{1}{1-\mu(E^1)}) dP(x)$, $E^1 = f_\circ(x)$, $E^2 = f(x)$ при некотором фиксированном значении $x \in D_X$.

Утверждение 3.8. Пусть стратегия природы c_k распознавания k образов порождена функцией f^* такой, что $f^*(x) = E_y^i$ при $x \in E_X^i$, тогда вероятность ошибки распознавания правилом f таким, что $f(x) = \omega_i$, $\omega_i \in E_y^i$, при $x \in E_X^i$, есть величина $P_f = 1 - \sum_{i=1}^{k} \frac{1}{k\mu(E_y^i)} p_x^i p_{y/x}^i$.

Следствие 3.8.1. $P_f + F(c_k, f^*) = 1 + \sum_{i=1}^{k} p_x^i \left[p_{y/x}^i \left(1 - \frac{1}{k\mu(E_y^i)} \right) - \mu(E_y^i) \right]$.

Утверждение 3.9. Множество всевозможных стратегий можно упорядочить по сложности, т.е. $L_{\tau^1}(1) \subset L_{\tau^2}(2) \subset \ldots \subset L_{\tau^s}(s) \subset \ldots \subset L_{\tau^M}(M) \subset \ldots \subset L^\circ$, причем $\tau^{s+1} \leq \tau^s$, где $M_{L(s)} = s$ - сложность класса стратегий природы, τ^s - допустимый уровень ошибки класса $L(s)$.

Предлагаемый метод $Q(v_N)$ построения выборочной решающей функции $\bar{f}$ основывается на максимизации эмпирического функционала качества $\overline{F}(\bar{f}) = \sum_{t=1}^{M'} \bar{p}_x^t (\bar{p}_{y/x}^t - \bar{\mu}_y^t)$, где $\bar{p}_x^t = \frac{N(\tilde{E}_X^t)}{N(D_X)} = \frac{N^t}{N}$, $\bar{p}_{y/x} = \frac{N(\hat{E}_y^t)}{N(\tilde{E}_X^t)} = \frac{\hat{N}^t}{N^t}$, $\bar{\mu}_y = \mu(\hat{E}_y)$, $N(\text{'*'})$ - число выборочных точек, попавших или образующих соответствующую область "*", $\bar{f} \sim < \alpha, r(\alpha) >$, $\alpha = \{\tilde{E}_X^1, \ldots, \tilde{E}_X^{M'}\} \in \Psi_{M'}$, $r(\alpha) = \{\hat{E}_y^1, \ldots, \hat{E}_y^{M'}\} \in R_{M'}$ Наилучшей выборочной решающей функцией является функция $\bar{f}^* = \arg \max_{\alpha \in \Psi_{M'}} \max_{r(\alpha) \in R_{M'}} \overline{F}(\bar{f})$. Для решения данной экстремальной задачи применяется

алгоритм MLRP последовательного увеличения ветвей дерева. Производится разветвление той вершины построенного дерева, для которой происходит максимальное увеличение значения критерия $\overline{F}(\bar{f})$, до тех пор пока вершина является делимой либо $\overline{F}(\bar{f}) \geq F^*$. Вершина дерева является неделимой, если 1) число конечных вершин $M' = M^*$ либо 2) $\hat{N}^t \leq N^*$. Критерий и параметры F^*, M^*, N^* определяют метод построения выборочной решающей функции.

Для оценивания качества предложенного метода ПСРП было проведено статистическое моделирование. Оценивалось математическое ожидание функционала качества: $m_F(c) = \mathbf{E}_{V_N} F(c, \bar{f})$ при фиксированной стратегии природы. Кроме того, с целью оценивания качества решения эмпирическим способом было исследовано смещение усредненного эмпирического функционала качества при фиксированной стратегии природы: $\varepsilon_N(c) = \mathbf{E}_{V_N} F(c, \bar{f}) - \mathbf{E}_{V_N} \overline{F}(\bar{f})$ и максимальное смещение математического ожидания эмпирического функционала качества при фиксированном значении эмпирического функционала: $\varepsilon_N^*(c) = \sup_{c:\overline{F}(\bar{f})=F_o} \varepsilon_N(c)$ для некоторых параметрических классов стратегий природы. Исследования проводились при фиксированной стратегии природы сложности M, сложности M' решающих функций, построенных алгоритмом MLRP по выборке объема *N*. Параметры *n, m* (размерности областей D_X и D_Y) и количество переменных определенного типа рассматривались в постановке задачи в целом и определяли сложность стратегии природы и решающей функции. Для моделирования параметров стратегий природы был разработан алгоритм GenMLRP. Генерирование стратегий природы осуществлялось в соответствии с введенным *определением 3.2*, где параметры задаются случайно в фиксированном диапазоне. При параметрах, определяющих равномерное распределение на всей области D_Y, значения функционала качества отражены в доказанных следствиях (*следствие 3.6.2, 3.6.3, 3.6.4*).

Результаты проведенных исследований в диссертации представлены в таблицах и графиках. Продемонстрируем некоторые из них. Например, для задачи прогнозирования *m* (*m*=1,2,3,4,5) непрерывных (*Cnt*) переменных по одной непрерывной переменной и стратегии природы, заданной параметрами *M*=1, $\min \mu(E_y^t) = 1$, $p_{y/x}^t = 1$, математическим моделированием получены зависимости $\bar{\varepsilon}_N(c)$ от размерности пространства прогнозируемых переменных и сложности решающей функции M'=1,2,3,4,5 при *N=20* (рис. 1).

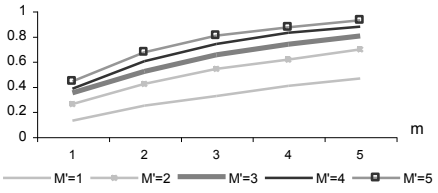

Рис. 1. Зависимость $\varepsilon_N(c)$ от размерности *m* при *N=20*.

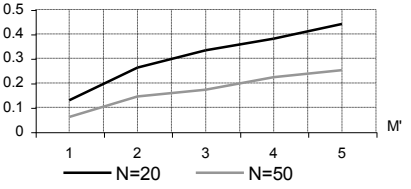

Рис. 2. Зависимость $\varepsilon_N(c)$ от сложности класса решающих функций при *m*=1.

На рис. 2 приводится график зависимости величины $\bar{\varepsilon}_N(c)$ от сложности класса решающих функций M' для объемов выборки 20 и 50 при m=1.

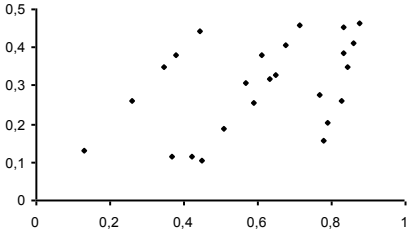

Рис. 3. Зависимость $\varepsilon_N(c)$ от $F_\circ$ при N=20

В следующем примере для непрерывного случая задачи прогнозирования одной переменной рассматриваются различные по параметрам $\mu(E_y^t)$ и $p_{y/x}^t$ стратегии природы сложности M=1. На графике (рис. 3) каждой точке соответствует значение оценки смещения $\bar{\varepsilon}_N(c)$ при фиксированном значении эмпирического функционала качества решающих функций. Рассматривались сложности $M' = 1,2,3,4,5$ и объем выборки равный 20. Полученные результаты дают возможность по значению эмпирического функционала качества предварительно оценить максимальное смещение $\bar{\varepsilon}_N^*(c)$. Результаты моделирования, приведенные в работе, демонстрируют на сколько и как изменяются исследуемые величины $\bar{\varepsilon}_N(c)$, $\bar{\varepsilon}_N^*(c)$, $\overline{m}_F(c)$ в зависимости от параметров M, M', N, n, m. Половина длины 95-ти процентного доверительного интервала для оцениваемых параметров имеет следующий порядок: 0.024 при N=8; 0.011 при N=20; 0.008 при N=50.

В §8 рассматривается дополнительная процедура учета эффекта влияния типа переменной на качество прогноза. Отметим, что предложенному критерию качества может удовлетворять не одна, а несколько подобластей с одинаковым значением меры $\hat{\mu}^t$ и с одинаковым числом точек $\hat{N}^t$, образующих данную подобласть, но с различным (упорядоченным или неупорядоченным) набором значений переменных. Процедура предпочтения одной области другой заключается в том, что менее вероятная подобласть в предположении равномерного распределения на всей области D_Y в большей степени претендует на "закономерность". Получено аналитическое и алгоритмическое представление вероятностей $P(\hat{E}_Y^t)$ в зависимости от типа переменной, меры подобласти и объема выборки, образовавшей её. На следующем графике представлены результаты различия этих вероятностей.

По оси ординат представлена сумма по всевозможным мерам абсолютной разности вероятностей образования "оболочек" для неупорядоченной и упорядоченной переменной

$$\Delta P = \sum_{i=1}^{L} |P_1(\mu^t, L, N^t) - P_2(\mu^t, L, N^t)|,$$

$\mu^t = i/L$; по оси абсцисс - отношение числа значений, принимаемых переменной, к числу точек, образующих подобласть

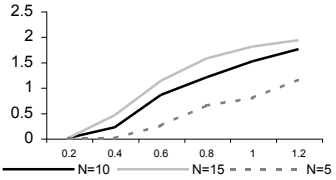

Рис. 4. Зависимость ΔP от L_j/N_t.

фиксированной меры. Результаты (рис. 4) показывают, что при достаточно больших значениях N^t и заданном L_j (уже при $\frac{L_j}{N^t} < 0.6$) эта разница незначительна и близка к нулю. При сравнимых значениях L_j и N^t или малых N^t различие существенно.

Четвертая глава посвящена демонстрации решения прикладных задач из области медицины и гидрологии, раскрывается их актуальность.

В первом параграфе рассматривается задача, поставленная научными сотрудниками института клинической и экспериментальной медицины, которая состоит в определении возможности применения унифицированного метода «рискометрии» в анализе взаимосвязи вероятности патологии с гелиогеофизическими характеристиками среды в пренатальный период жизни человека. На основе компьютерной базы исследовательских данных была сформирована выборка лиц обоего пола в количестве 1556 человек в возрасте от 19 до 67 лет, у которых был определен риск по каждому из 11 патологических синдромов. На каждого вошедшего в указанную выборку, с помощью компьютерной программы «Cosmic – V.01», была получена информация о гелиогеофизической обстановке усредненно на каждую из 40 недель, предшествующих дате рождения. В классе логических решающих функций выявлены логические закономерности взаимосвязи между величинами показателей солнечной активности, обобщенной характеристики напряженности магнитного поля Земли на различных сроках пренатального развития и уровнями риска патологических синдромов на момент исследования. Целевыми данными явились признаки, представляющие собой количественный показатель вероятности риска патологических синдромов: АГ- артериальная гипертония, ИБС- болезнь сердца, ЖКТ- нарушение деятельности органов желудочно-кишечного тракта, ПЕЧ- печени, ЛЕГ- органов дыхания, ЭНД- эндокринной системы, ИММ- иммунной системы, РЕН- почек, НРВ- неврологических заболеваний, ПСИ- психические, полученные с помощью математической программы АСКОРС в лаборатории клинической физиологии ИКЭМ. Для каждого синдрома полученные закономерности (порядка 14) объединяются в дерево решений, которое легко интерпретируется на языке близком к обычному языку высказываний. На рис. 5 приведено дерево решений для ИБС величины риска синдрома в группах, характеризующихся годом рождения (Гр) и различной величиной солнечной активности (Оисп) и геомагнитной индукции (Игв). Качество решения было оценено вероятностью ошибки на контроле и равнялось 0.3, что являлось вполне удовлетворительным результатом.

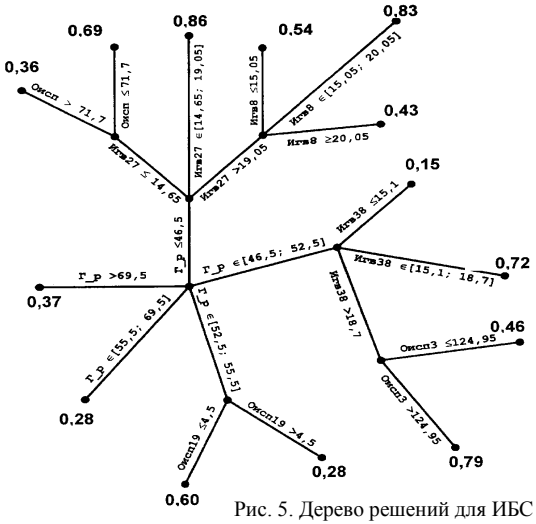

Рис. 5. Дерево решений для ИБС

Во втором параграфе рассматривается решение задачи прогнозирования водосбора воды, проходящей через русло реки Обь, среднемесячной температуры и осад-

ков по данным метеорологических наблюдений в контрольной точке за 86 лет (с 1913 по 2000 год). Были обработаны данные ежемесячных замеров за ноябрь, декабрь, январь и апрель месяцы в контрольной точке г. Колпашево с целью построения решающего правила предсказания трех переменных в апреле по трем переменным за три месяца (ноябрь, декабрь, январь). Решающая функция прогноза была построена по обучающей выборке объема 76. Оценка критерия качества (оценка вероятности принятия верного решения по правилу $\bar{f}$) получена по контрольной выборке объема 10 (последние десять лет) и равна 0.8, что является вполне удовлетворительным результатом. При последовательном построении дерева решений были выделены четыре значимых признака из девяти, по которым и было проведено разбиение. В соответствии с построенным правилом оказалось, что наибольшее влияние на качество прогноза оказывают среднемесячная температура в ноябре (X1) и январе (X3), осадки в ноябре (X4) и водосбор в ноябре (X7).

Основные результаты работы состоят в следующем:

1. Разработан способ оценивания качества метода прогнозирования системы разнотипных переменных (ПСРП).
2. Предложен метод построения логико-вероятностной модели прогнозирования системы разнотипных переменных, основанный на предложенном критерии качества.
3. Получены зависимости, позволяющие определить влияние типа переменной (с упорядоченным и неупорядоченным набором значений) на качество решения при ПСРП в условиях малой выборки.
4. Проведены исследования зависимости качества метода ПСРП от сложности распределения, сложности класса решающих функций и объема выборки.
5. Для порогового метода построения кусочно-постоянных решающих функций (одномерный случай) при заданном классе распределений получена нижняя оценка его качества в зависимости от сложности класса решающих функций (число областей разбиения) и объема выборки.
6. Предложен критерий обнаружения значимого подмножества переменных МНК-метода построения линейной регрессионной функции.
7. На модельных примерах и прикладных задачах продемонстрирована эффективность предложенных методов.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1. Ступина Т.А. О соотношении прозрачности воды и концентрации фитопланктона в Байкале.// Сб. статей «математические проблемы экологии», Новосибирск, 1994 – с. 125-128.
2. G.S. Lbov, T.A. Stupina. Некоторые вопросы устойчивости выборочных решающих функций.// Pattern Recognition and Image Analysis, Vol 9, N 3, 1999 – pp. 408-415.
3. Лбов Г.С., Ступина Т.А. О статистической устойчивости решающих функций в задачах распознавания и регрессионного анализа.// ДАН, 1999, том 368, N1. – с. 31-34.
4. Ступина Т.А. Задача предсказания многомерной переменной.// Доклады IX Всероссийской конференции «Математические методы распознавания образов». РАН ВЦ, 1999 – с. 67-69.
5. В.П. Казначеев, Я.В. Поляков, А.В. Трофимов, Г.С. Лбов, Т.А. Ступина и др. Гелиогеофизические факторы среды при пренатальном развитии в вероятностной модели прогноза здоровья человека.// Весник МНИКА, выпуск N6, 1999, Н-ск. – с. 37-43.
6. G.S. Lbov, T.A. Stupina. The influence of type of objective variabe on quality of prediction.// Proccedings of the Sixth International Conference, Minsk, 2001 – pp. 250-253.
7. Лбов Г.С., Ступина Т.А. О критерии качества решающей функции предсказания многомерной переменной.// Доклады X Всероссийской конференции ММРО, Москва 2001 – с. 138-141.

8. Ступина Т.А. О критерии значимости переменных в линейном регрессионном анализе в условиях малых выборок.// VI Международная конференция «Современные методы математического моделирования природных и антропогенных катастроф», Красноярск 2001 – с. 270-274.
9. Лбов Г.С., Ступина Т.А. О критерии качества решающей функции при прогнозировании многомерной переменной.// Таврический вестник информатики и математики. Изд-во НАН Украины, 2002 – с. 172-179.
10. Лбов Г.С., Ступина Т.А. Построение функции прогноза в многомерном разнотипном пространстве.// Труды I международной конференции «Информационные системы и технологии (IST'2002)», Минск, 2002 – с. 253-254.
11. G.S. Lbov, T.A. Stupina. To question of statistical stability of sampling decision function of prediction multidimensional variable.// Proceeding of the seven international conference, (PRIP'2003), Minsk, Vol 2. – с. 57-61.
12. Лбов Г.С., Ступина Т.А., Полякова Г.Л. Метод обнаружения закономерностей для прогнозирования многомерной разнотипной переменной.// Труды Всероссийской конференции «Математические и информационные технологии в энергетике, экономике, экологии», Иркутск, 2003 – с.199-203.
13. G.S. Lbov, T.A. Stupina, V.B. Berikov, A.A. Vikent'ev. On statistical stability of sample decision function in pattern recognition and prediction.// The 6-th german-russion workshop "Pattern recognition and Image Understanding", Altai, 2003 – pp. 46-49.
14. Лбов Г.С., Ступина Т.А. Исследование зависимости критерия качества прогнозирования многомерной переменной от объема выборки и сложности решающей функции.// Труды XI Всероссийской конференции ММРО-03, Москва, 2003 – с. 127-129.
15. G.S. Lbov, T.A. Stupina. Statistical Stability of Sampling Decision Functions in Recognition and Prediction Problems.// "Pattern Recognition and Image Analysis", Vol. 14, No 2'2004 – pp. 231-236.
16. Г.С. Лбов, Т.А. Ступина. Исследование эффективности метода прогнозирования многомерной переменной.// Таврический вестник информатики и математики. Изд-во НАН Украины 2004, № 1 – с. 117-122.
17. T.A. Stupina. The Properties of Risk Function In Heterogeneous Multivariate Prediction.// Proceeding of the 8 international conference, (PRIP'2005), Minsk, Vol 1 – pp. 80-83.
18. Лбов Г.С., Бериков В.Б., Герасимов М.А., Ступина Т.А. Анализ многомерных разнотипных временных рядов для прогнозирования экстремальных гидрологических ситуаций.// II-Международная конференция «Фундаментальные проблемы изучения и использования воды и водных ресурсов», Иркутск, 2005 – с. 235-236.
19. Ступина Т.А. Оценка смещения функционала качества в задаче прогнозирования многомерной разнотипной переменной.// Доклады Всероссийской конференции "Математические методы распознавания образов (ММРО-12)", Москва, 20–26 ноября 2005 – с. 209-212.
20. G.S. Lbov, T.A. Stupina. Application of the multivariate prediction method to time series.// International Journal ITHEA, Vol 13, No 3'2006 – pp. 278-285.
21. T.A. Stupina. Recognition of the Heterogeneous Multivariate Variable.// Proceeding of the international conference, 2006 (KDS'2006), Varna (Bulgaria), Vol 1 – pp. 199-202.