РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ

2004, том 40, № 1

УДК 519.7

В. А. Лапко
*(Красноярск)*

# СИНТЕЗ И АНАЛИЗ ГИБРИДНЫХ МОДЕЛЕЙ СТОХАСТИЧЕСКИХ ЗАВИСИМОСТЕЙ ПРИ НАЛИЧИИ ИХ ЧАСТНОГО ОПИСАНИЯ*

Рассматриваются гибридные модели стохастических зависимостей при наличии априорных сведений об их описании в неполном пространстве контролируемых признаков. Исследуются асимптотические свойства предлагаемых моделей, которые сопоставляются с результатами вычислительного эксперимента.

**Введение.** Традиционные методы аппроксимации параметрического и непараметрического типа [1] в основном ориентируются либо на априорную информацию о виде восстанавливаемой зависимости, либо на экспериментальные данные о ее локальном поведении, что ограничивает их эффективность. Традиционные гибридные модели [2] сочетают в одном решающем правиле преимущества параметрических и непараметрических аппроксимаций. При этом единое решающее правило образуют параметрическая модель восстанавливаемой зависимости и непараметрическая оценка функции невязки, которые строятся в одном и том же пространстве переменных. Особенность рассматриваемых модификаций гибридных моделей состоит в том, что искомая зависимость $y = \varphi(x)\ \forall\ x \in R^k$ представлена обучающей выборкой $V = (x^i,\ y^i,\ i = \overline{1, n})$ и имеется ее частное описание $\bar{y}_1 = F(x_1, \alpha)$ в ограниченном пространстве контролируемых признаков $x_1 \in R^{k1},\ k1 < k$. Для максимального учета априорных сведений предлагается на основе принципов гибридного моделирования объединить в одном решающем правиле частное описание $F(x_1, \alpha)$ и информацию об искомой зависимости, содержащейся в обучающей выборке $V$.

Актуальность рассматриваемой проблемы подтверждается перспективностью применения методики ее решения в процессе исследования статических объектов при наличии их частных описаний $\bar{y}_1 = F(x_1, \alpha)$, где $x_1 = (x_{1v},\ v = \overline{1, k1})$, $y$ – входные и выходные переменные соответственно. При появлении возможности контроля дополнительного набора компонент входных переменных изучаемого объекта $x_2 = (x_{2v},\ v = \overline{1, k2})$, оказывающих существен-

* Работа выполнена при поддержке гранта Президента Российской Федерации (№ МК-143.2003.01).

ное влияние на изменение выходной переменной $y$, возникает необходимость построения модели зависимости $y = F(x_1, x_2)$ на основании априорной информации $\bar{y}_1 = F(x_1, \alpha)$ и экспериментальных данных $V = (x_v^i,\ y^i,$ $v = \overline{1, k},\ i = \overline{1, n})$.

**Постановка задачи.** Пусть для искомой однозначной зависимости

$$y = \varphi(x)\ \forall\ x \in R^k \tag{1}$$

известно ее частное описание относительно некоторого ограниченного набора признаков:

$$\bar{y}_1 = F(x_1, \alpha)\ \forall\ x_1 \in R^{k1}, \quad k1 < k,$$

и выборка $V = (x_v^i, y^i,\ v = \overline{1, k},\ i = \overline{1, n})$ экспериментальных данных, составленная из статистически независимых значений переменной $(x, y)$ исследуемой зависимости (1).

Задача состоит в построении модифицированной гибридной модели $\bar{\bar{y}}(x)$ зависимости (1), совмещающей в одном решающем правиле всю имеющуюся априорную информацию.

**Синтез модифицированной гибридной модели с учетом частного описания.** На первом этапе синтеза структуры модифицированной гибридной модели при использовании статистической выборки $V_1 = (x_v^i, y^i,\ v = \overline{1, k1},$ $i = \overline{1, n})$ проводится идентификация параметров $\alpha$ модели $\bar{y}_1 = F(x_1, \alpha)$.

Далее формируется выборка

$$V_2 = (x_v^i,\ q(x_v^i, v = \overline{k1+1, k}),\ v = \overline{k1+1, k},\ i = \overline{1, n}),$$

составленная из значений функции невязок

$$q(\bar{x}_1^i = (x_v^i, v = \overline{k1+1, k})) = y^i - F(x_1^i, \bar{\alpha})$$

между экспериментальными данными и параметрической моделью $\bar{y}_1 = F(x_1, \bar{\alpha})$ в пространстве $x_v$, $v = \overline{k1+1, k}$, где $\bar{\alpha}$ – оценки параметров $\alpha$ модели $F(x_1, \alpha)$.

Для восстановления функции невязок по выборке $V_2$ воспользуемся непараметрической регрессией

$$\bar{q}(\bar{x}_1) = \sum_{i=1}^{n} q(\bar{x}_1^i)\beta_i(\bar{x}_1), \tag{2}$$

где

$$\beta_i(\bar{x}_1) = \frac{\prod\limits_{v=k1+1}^{k} \Phi\left(\dfrac{x_v - x_v^i}{c_v}\right)}{\sum\limits_{i=1}^{n} \prod\limits_{v=k1+1}^{k} \Phi\left(\dfrac{x_v - x_v^i}{c_v}\right)}.$$

Здесь $\Phi(u) \geq 0$ – ядерная функция, удовлетворяющая свойствам

$$\Phi(u) = \Phi(-u), \qquad \int \Phi(u)du = 1, \qquad \int u^m \Phi(u)du < \infty.$$

Тогда гибридная модель стохастической зависимости (1) с учетом ее частного описания $F(x_1, \bar{\alpha})$ представляется статистикой

$$\bar{\bar{y}}(x) = \bar{\varphi}(x_1, \bar{x}_1) = F(x_1, \bar{\alpha}) + \bar{q}(\bar{x}_1). \tag{3}$$

Ближайшим аналогом данной модели являются гибридные аппроксимации [2], в которых параметрическая модель и оценка функции невязки определяются в пространстве полного набора компонент вектора $x = (x_1, \bar{x}_1)$.

**Асимптотические свойства модели.** Свойства гибридной модели (3) определяются следующим утверждением.

**Теорема.** Пусть

1) восстанавливаемая зависимость $\varphi(x)$ представима суммой однозначных функций $\varphi(x) = \varphi_1(x_1) + \varphi_2(\bar{x}_1)$;

2) функции $\varphi_1(x_1)$, $\varphi_2(\bar{x}_1)$ и плотности вероятности $p(x)$, $p(\bar{x}_1)$, $p(x_1)$ ограничены вместе со своими производными до второго порядка включительно;

3) $\Phi(u)$ относится к классу ограниченных, положительных, симметричных и нормированных функций;

4) последовательность параметров $c(n) \geq 0$ ядерных функций $\Phi(\cdot)$ такова, что при $n \to \infty$ значения $c(n) \to 0$, а $nc \to \infty$.

Тогда модифицированная гибридная модель (3) обладает свойствами асимптотической несмещенности и состоятельности.

Доказательство. Вычислим математическое ожидание смещения при известном законе распределения $\bar{x}_1 \in R^1$:

$$M(\varphi(x_1, \bar{x}_1) - \bar{\varphi}(x_1, \bar{x}_1)) = M(\varphi(x_1, \bar{x}_1) - F(x_1, \alpha) - \bar{q}(\bar{x}_1)) =$$

$$= \varphi(x_1, \bar{x}_1) - MF(x_1, \alpha) - M\bar{q}(\bar{x}_1), \tag{4}$$

где

$$M\bar{q}(\bar{x}_1) = \frac{1}{ncp(\bar{x}_1)} \sum_{i=1}^{n} M\left( (y^i - F(x_1^i, \alpha))\Phi\left( \frac{\bar{x}_1 - \bar{x}_1^i}{c} \right) \right) =$$

$$= \frac{1}{cp(\bar{x}_1)} \int\int\int (y - F(t_1, \alpha))\Phi\left( \frac{\bar{x}_1 - t_2}{c} \right) p(y, t_1, t_2) dy dt_1 dt_2.$$

Учитывая, что оптимальным решающим правилом в среднеквадратическом смысле является условное математическое ожидание

$$\int y p(y/(t_1, t_2)) dy = \varphi(t_1, t_2),$$

получим

$$M\bar{q}(\bar{x}_1)=\frac{1}{cp(\bar{x}_1)}\int\int\varphi(t_1,t_2)p(t_1/t_2)dt_1\Phi\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_2)dt_2=$$

$$=\frac{1}{cp(\bar{x}_1)}\int\varphi_2(t_2)\Phi\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_2)dt_2.$$

Проведем замену переменных $(\bar{x}_1-t_2)/c=u$ и разложим функции $\varphi_2(\bar{x}_1-cu)$, $p(\bar{x}_1-cu)$ в ряд Тейлора в окрестности точки $\bar{x}_1$.

Тогда

$$M\bar{q}(\bar{x}_1)=p^{-1}(\bar{x}_1)\int\Phi(u)\left(\varphi_2(\bar{x}_1)-cu\varphi_2^{(1)}(\bar{x}_1)+\frac{c^2u^2}{2}\varphi_2^{(2)}(\bar{x}_1)+\ldots\right)\times$$

$$\times\left(p(\bar{x}_1)-cup^{(1)}(\bar{x}_1)+\frac{c^2u^2}{2}p^{(2)}(\bar{x}_1)+\ldots\right)du.$$

Учитывая справедливость соотношения $\int u\Phi(u)du=0$ и принимая $\int u^2\Phi(u)du=1$, при достаточно больших $n$ имеем

$$M\bar{q}(\bar{x}_1)\sim\varphi_2(\bar{x}_1)+c^2\left[\frac{p^{(2)}(\bar{x}_1)\varphi_2(\bar{x}_1)}{2p(\bar{x}_1)}+\frac{p^{(1)}(\bar{x}_1)\varphi_2^{(1)}(\bar{x}_1)}{p(\bar{x}_1)}+\frac{\varphi_2^{(2)}(\bar{x}_1)}{2}\right]+$$

$$+\frac{c^4\varphi_2^{(2)}(\bar{x}_1)p^{(2)}(\bar{x}_1)}{4p(\bar{x}_1)}+O(c^6). \tag{5}$$

При $c(n)\to 0$ с ростом $n\to\infty$ выражение $M\bar{q}(\bar{x}_1)$ стремится к $\varphi_2(\bar{x}_1)$. Так как в соответствии с условиями теоремы $\varphi(x)=\varphi_1(x_1)+\varphi_2(\bar{x}_1)$, то смещение (4) определяется свойствами частного описания $F(x_1,\alpha)$, т. е. $M(\varphi_1(x_1)-F(x_1,\alpha))$. Поэтому, если модель $F(x_1,\alpha)$ обладает свойствами асимптотической несмещенности, оно присуще и модифицированной гибридной модели (3).

Для доказательства сходимости статистики (3) в среднеквадратическом выполним очевидные преобразования:

$$M(\varphi(x_1,\bar{x}_1)-\bar{\varphi}(x_1,\bar{x}_1))^2=M[(\varphi_1(x_1)-F(x_1,\alpha))+(\varphi_2(\bar{x}_1)-\bar{q}(\bar{x}_1))]^2=$$

$$=M(\varphi_1(x_1)-F(x_1,\alpha))^2+2M[(\varphi_1(x_1)-F(x_1,\alpha))(\varphi_2(\bar{x}_1)-\bar{q}(\bar{x}_1))]+$$

$$+M(\varphi_2(\bar{x}_1)-\bar{q}(\bar{x}_1))^2.$$

Применяя ко второму слагаемому неравенство Шварца, получим

$$M(\varphi(x)-\overline{\varphi}(x))^2<$$

$$<\left[(M(\varphi_1(x_1)-F(x_1,\alpha))^2)^{1/2}+(M(\varphi_2(\bar{x}_1)-\bar{q}(\bar{x}_1))^2)^{1/2}\right]^2. \tag{6}$$

Оценим асимптотическое выражение

$$M(\varphi_2(\bar{x}_1)-\bar{q}(\bar{x}_1))^2=\varphi_2^2(\bar{x}_1)-2\varphi_2(\bar{x}_1)M\bar{q}(\bar{x}_1)+M\bar{q}^2(\bar{x}_1),$$

где

$$M\bar{q}^2(\bar{x}_1)=\frac{1}{n^2c^2p^2(\bar{x}_1)}\left[\sum_{i=1}^{n}M\left((y^i-F(x_1^i,\alpha))^2\Phi^2\left(\frac{\bar{x}_1-\bar{x}_1^i}{c}\right)\right)+\right.$$

$$\left.+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}M\left[(y^i-F(x_1^i,\alpha))\Phi\left(\frac{\bar{x}_1-\bar{x}_1^i}{c}\right)(y^j-F(x_1^j,\alpha))\Phi\left(\frac{\bar{x}_1-\bar{x}_1^j}{c}\right)\right]\right]=$$

$$=\frac{1}{n^2c^2p^2(\bar{x}_1)}\left[nM\left((y-F(t_1,\alpha))^2\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)\right)+\right.$$

$$\left.+n(n-1)\left(M\left((y-F(t_1,\alpha))\Phi\left(\frac{\bar{x}_1-t_2}{c}\right)\right)\right)^2\right]. \tag{7}$$

Пренебрегая величинами степени малости менее $O(c^4)$ в выражении (5), определим второе слагаемое соотношения (7):

$$\varphi_2^2(\bar{x}_1)+c^2\varphi_2(\bar{x}_1)B_1(\bar{x}_1)+c^4B_1^2(\bar{x}_1), \tag{8}$$

где

$$B_1(\bar{x}_1)=\frac{p_2^{(2)}(\bar{x}_1)\varphi_2(\bar{x}_1)}{p(\bar{x}_1)}+\frac{2\varphi_2^{(1)}(\bar{x}_1)p^{(1)}(\bar{x}_1)}{p(\bar{x}_1)}+\varphi_2^{(2)}(\bar{x}_1).$$

Оценим первое слагаемое выражения (7):

$$\frac{1}{nc^2p^2(\bar{x}_1)}M\left((y-F(t_1,\alpha))^2\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)\right)=$$

$$=\frac{1}{nc^2p^2(\bar{x}_1)}\int\int(\varphi(t_1,t_2)-F(t_1,\alpha))^2\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_1,t_2)dt_1dt_2=$$

$$=\frac{1}{nc^2p^2(\bar{x}_1)}\int\int[(\varphi_1(t_1)-F(t_1,\alpha))+\varphi_2(t_2)]^2\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_1,t_2)dt_1dt_2. \tag{9}$$

Обозначим

$$\beta_1=\max_{t_1}\int(\varphi_1(t_1)-F(t_1,\alpha))^2\,p(t_1/t_2)dt_1,$$

$$\beta_2=\max_{t_2}\int|\varphi_1(t_1)-F(t_1,\alpha)|p(t_1/t_2)dt_1.$$

Тогда ограничение выражения (9) принимает вид

$$\frac{1}{nc^2p^2(\bar{x}_1)}\left[\beta_1\int\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_2)dt_2+2\beta_2\int\varphi_2(t_2)\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_2)dt_2+\right.$$

$$\left.+\int\varphi_2(t_2)\Phi^2\left(\frac{\bar{x}_1-t_2}{c}\right)p(t_2)dt_2\right].$$

Проведем замену переменных $(\bar{x}_1-t_2)/c=u$ и в результате преобразований легко получим асимптотическую оценку ограничения первого слагаемого выражения (7):

$$\frac{\|\Phi(u)\|^2}{ncp(\bar{x}_1)}(\beta_1+2\beta_2\varphi_2(\bar{x}_1)+\varphi_2^2(\bar{x}_1))+O(c/n). \tag{10}$$

С учетом (5), (8) и (10) запишем окончательное выражение среднеквадратического отклонения

$$M(\varphi_2(\bar{x}_1)-\bar{q}(\bar{x}_1))^2<\frac{\|\Phi(u)\|^2}{ncp(\bar{x}_1)}(\beta_1+2\beta_2\varphi_2(\bar{x}_1)+\varphi_2^2(\bar{x}_1))+$$

$$+c^4\left(B_1^2(\bar{x}_1)-\frac{\varphi_2^{(2)}(\bar{x}_1)p^{(2)}(\bar{x}_1)\varphi_2(\bar{x}_1)}{2p(\bar{x}_1)}\right).$$

Из анализа (6) при $c\to 0$, $nc\to\infty\ \forall\ n\to\infty$ следует сходимость в среднеквадратическом модифицированной гибридной модели (3), если таким же свойством обладает модель $F(x_1,\alpha)$. Необходимо отметить зависимость главного члена дисперсии непараметрической оценки функции невязки $q(\bar{x}_1)$ от среднеквадратического отклонения и смещения модели $F(x_1,\alpha)$.

**Анализ свойств модели (3) при конечных объемах обучающих выборок.** В качестве искомой зависимости (1) использовался полином вида

$$y(x)=\sum_{j=1}^{k}[-4x_j^2+4x_j]. \tag{11}$$

При формировании исходной выборки $V=(x_v^i,\, y^i,\, v=\overline{1,k},\, i=\overline{1,n})$ значения функции зашумлялись аддитивной помехой

$$y^i+2y^i(\varepsilon-0{,}5)r,$$

где $\varepsilon$ – случайная величина с равномерным законом распределения на интервале [0, 1], а $r\cdot 100\,\%$ – уровень помех.

Частные сведения о восстанавливаемой зависимости (11) представлялись одним из полиномов:

$$F(x,\overline{\alpha})=\sum_{j=1}^{k1}[-4x_j^2+4x_j], \tag{12}$$

$$F(x,\overline{\alpha})=\sum_{j=1}^{k1}[-5x_j^2+5x_j], \tag{13}$$

где $k1$ – размер вектора $x(1)=(x_1,\ldots,x_{k1})$.

Исследовались зависимости статистических критериев аппроксимации функции (11)

$$W_1=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y^i-\bar{\bar{y}}(x^i)}{y^i}\right|, \tag{14}$$

$$W_2=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y(x^i)-\bar{\bar{y}}(x^i)}{y(x^i)}\right| \tag{15}$$

модифицированной гибридной моделью $\bar{\bar{y}}(x)$ от объема $n$ экспериментальных данных $V=(x_v^i,\, y^i,\, v=\overline{1,k},\, i=\overline{1,n})$, уровня помех $r\cdot 100\,\%$ и размера $k2=k-k1$ пространства признаков $x(2)$. Для критерия (15) значения $y(x^i)$ вычислялись с использованием полинома (11), который применяется для генерации обучающей выборки.

В ходе проведения вычислительного эксперимента установлено, что доступная для контроля ошибка типа (14) слабо чувствительна по сравнению с критерием (15) к изменению аппроксимационных свойств непараметрической регрессии и модифицированной гибридной модели с учетом частного описания (3). Данный факт особо проявляется при исследовании влияния уровня помех $r\cdot 100\,\%$ на аппроксимационные свойства модели (3) (рис. 1, *а*).

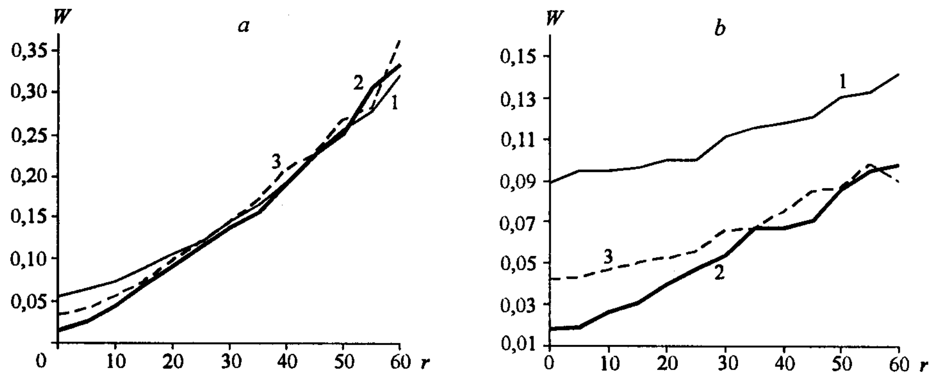


*Рис. 1.* Зависимости ошибок аппроксимации моделей типа (14) (*a*) и (15) (*b*) от уровня помех при $n = 500, k1 = 2, k2 = 2$: непараметрическая регрессия (кривые 1); гибридная модель (3) при отсутствии погрешности определения коэффициентов в частном описании (12) искомой зависимости (кривые 2); гибридная модель (3) при наличии 25 % погрешности (13) определения всех коэффициентов в частном описании (кривые 3)

Тенденции изменения критериев (14), (15) в зависимости от объема обучающей выборки $n$ при относительно малом уровне помех (10 %) и отношении $k1/k2 = 1$ размеров векторов признаков $x(1), x(2)$ близки (рис. 2). Тем самым обосновывается возможность использования эмпирического критерия (14) в рассмотренных условиях.

Искажение частных сведений о виде зависимости (11) приводит к некоторому ухудшению качества аппроксимации модифицированной гибридной модели (см. рис. 2, кривые 3), которая сохраняет преимущество перед непараметрической регрессией и является менее чувствительной к изменению объема обучающей выборки.

С увеличением уровня помех преимущество гибридной модели с учетом частного описания (3) сохраняется перед непараметрической регрессией (рис. 1, *b*), которое не вскрывается критерием (14) (см. рис. 1, *a*).

Снижение относительной значимости частных сведений за счет увеличения количества дополнительных признаков $k2$ приводит к ожидаемому снижению аппроксимационных свойств исследуемой модели (3), что отражено на рис. 3. Однако ее преимущество перед непараметрической регрессией сохраняется.

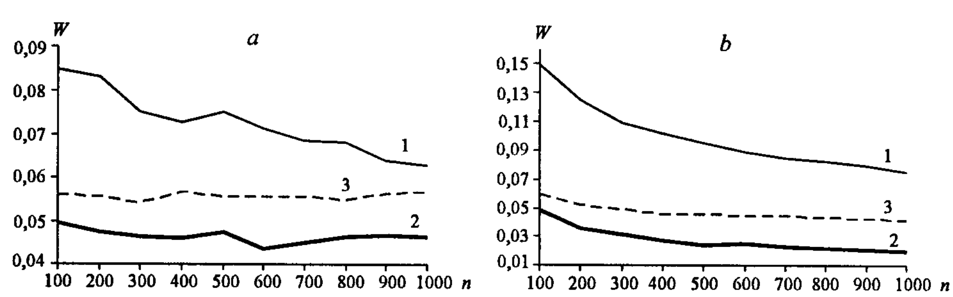


*Рис. 2.* Зависимости ошибок аппроксимации моделей типа (14) (*a*) и (15) (*b*) от объема выборки при $r = 10$ %, $k1 = 2, k2 = 2$. (Описание кривых такое же, как на рис. 1)

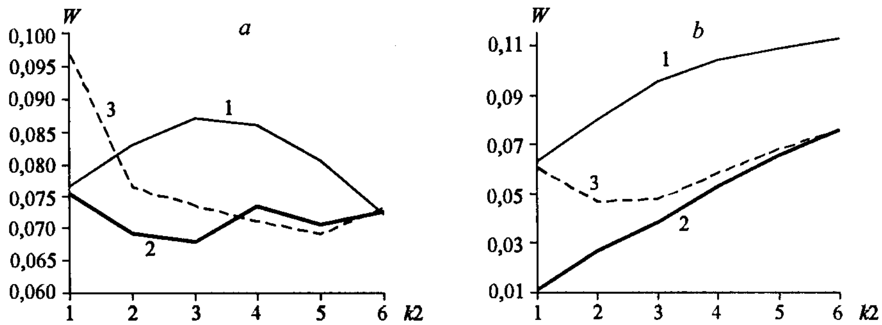


*Рис. 3.* Зависимости ошибок аппроксимации моделей типа (14) (*a*) и (15) (*b*) от количества признаков $k2$ при $n = 1000$, $k1 = 2$, $r = 15$ %. (Описание кривых такое же, как на рис. 1)

**Заключение.** Гибридные модели стохастических зависимостей с учетом их частного описания позволяют наиболее полно использовать априорную информацию в условиях расширения возможностей систем контроля изучаемых объектов. Установлены свойства асимптотической несмещенности и состоятельности предлагаемых моделей, что создает объективную основу для их сравнения с традиционными стохастическими аппроксимациями. С помощью метода статистического моделирования подтверждена эффективность гибридной модели с частным описанием по сравнению с непараметрической регрессией при различных условиях эксперимента.

## СПИСОК ЛИТЕРАТУРЫ

1. **Хардле В.** Прикладная непараметрическая регрессия. М.: Мир, 1993.
2. **Лапко А. В., Ченцов С. В., Крохов С. И., Фельдман Л. А.** Обучающиеся системы обработки информации и принятия решений. Новосибирск: Наука, 1996.

*Институт вычислительного моделирования СО РАН,*
*E-mail: lapko@ksc.krasn.ru*

*Поступила в редакцию*
*21 июня 2003 г.*