УДК 519.7

А. В. Лапко, В. А. Лапко

НЕПАРАМЕТРИЧЕСКАЯ РЕГРЕССИЯ В УСЛОВИЯХ ЧАСТИЧНЫХ СВЕДЕНИЙ О ВИДЕ ВОССТАНАВЛИВАЕМЫХ ЗАКОНОМЕРНОСТЕЙ

Предлагается непараметрическая модель стохастической зависимости, которая обеспечивает учёт априорных сведений о виде восстанавливаемых закономерностей. Исследуются её асимптотические свойства, проводится анализ результатов вычислительных экспериментов.

Ключевые слова: непараметрическая регрессия, стохастические зависимости, априорные сведения, асимптотические свойства.

Для наиболее полного учёта априорной информации о виде восстанавливаемых зависимостей и экспериментальных данных о её локальном поведении широко используются гибридные модели [1]. Традиционные гибридные модели сочетают в одном решающем правиле преимущество параметрических и непараметрических аппроксимаций. При этом единое решающее правило образуют параметрическая модель восстанавливаемой зависимости и корректирующая её функция непараметрического типа, которые строятся в одном и том же пространстве переменных. Полученные результаты были развиты для условий наличия частичных априорных сведений о виде восстанавливаемых зависимостей в ограниченном пространстве признаков [2]. Основная проблема применения гибридных моделей состоит в выборе вида корректирующей функции, которая является трудно формализуемой. Для её обхода предлагается использовать непараметрическую регрессию, синтез которой основан на обобщении априорной информации о виде восстанавливаемых зависимостей и экспериментальных данных об их локальном поведении.

Цель работы состоит в обосновании возможности учёта априорных сведений о виде восстанавливаемых закономерностей при синтезе непараметрических моделей стохастических зависимостей, основанных на оценках плотности вероятности типа Розенблатта–Парзена [3].

Синтез модифицированной непараметрической регрессии. Пусть об искомой однозначной зависимости

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

известно её описание

$$\bar{y}_1 = F(\bar{x}_1, \alpha) \quad \forall \quad \bar{x}_1 \in R^{k2}, \quad k2 < k$$

относительно некоторого ограниченного набора признаков из $x = (x_1, \bar{x}_1)$, $x_1 = (x_{1v}, v = \overline{1, k1})$, $k = k1 + k2$ и выборка $V = (x_v^i, v = \overline{1, k}, y^i, i = \overline{1, n})$ экспериментальных данных, составленная из статистически независимых значений переменных x, y исследуемой зависимости (1). Параметры α полинома $F(\bar{x}_1, \alpha)$ заданы.

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

На основании исходных экспериментальных данных сформируем промежуточную обучающую выборку

$$V^1 = (x_1^i, \bar{y}_1^i = F(\bar{x}_1^i, \alpha), y^i, i = \overline{1, n}).$$

Принимая в качестве оптимального решающего правила, в смысле минимума среднеквадратического отклонения, условное математическое ожидание $\varphi(x)$ [4], построим непараметрическую регрессию

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

где

$$\beta_i(x) = \frac{\prod_{v=1}^{k1} \Phi\left(\frac{x_{1v} - x_{1v}^i}{c_v}\right) \Phi\left(\frac{\bar{y}_1 - \bar{y}_1^i}{c}\right)}{\sum_{i=1}^{n} \prod_{v=1}^{k1} \Phi\left(\frac{x_{1v} - x_{1v}^i}{c_v}\right) \Phi\left(\frac{\bar{y}_1 - \bar{y}_1^i}{c}\right)}.$$

В статистике (2) ядерные функции $\Phi(u)$ удовлетворяют условиям H:

$$\Phi(u) = \Phi(-u), \; 0 \le \Phi(u) < \infty,$$
$$\int \Phi(u)\, du = 1, \; \int u^2 \Phi(u)\, du = 1,$$
$$\int u^m \Phi(u)\, du < \infty, \; 0 \le m < \infty;$$

$c = c(n)$, $c_v = c_v(n)$, $v = \overline{1, k1}$ – коэффициенты размытости ядерных функций, значения которых убывают с ростом объёма n обучающей выборки. Здесь и далее бесконечные пределы интегрирования опускаются.

При оценивании зависимости в ситуациях $x = (x_1, \bar{x}_1)$ сначала вычисляется $\bar{y}_1 = F(\bar{x}_1, \alpha)$, а затем по данным $(x_1, \bar{y}_1)$ в соответствии со статистикой (2) определяется значение $\bar{\bar{y}}(x)$.

Оптимизация модифицированной непараметрической регрессии (2) по коэффициентам размытости

ядерных функций c, c_v, $v=\overline{1,k1}$ осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки среднеквадратической ошибки аппроксимации искомой зависимости.

Асимптотические свойства непараметрической регрессии. Без существенной потери общности будем считать, что в частичном наборе признаков x_{1v}, $v=\overline{1,k1}$ их количество $k1 = 1$. В качестве ядерной функции примем функцию вида

$$\Phi(u)=\begin{cases}\frac{1}{2} & \forall \quad |u|<1\\ 0 & \forall \quad |u|\ge 1.\end{cases}$$

В этом случае непараметрическая регрессия (2) запишется как

$$\overline{\overline{y}}(x)=\frac{\sum_{i=1}^{n} y^i \Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\overline{y}_1-\overline{y}_1^i}{c}\right)}{\sum_{i=1}^{n}\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\overline{y}_1-\overline{y}_1^i}{c}\right)}. \qquad (3)$$

Тогда справедливо следующее утверждение.

Теорема. Пусть 1) частичные сведения $\overline{y}_1=F(\overline{x}_1,\alpha)$ восстанавливаемой зависимости (1) принадлежат к классу линейных полиномов; 2) функция $y=\varphi(x)$ и плотности вероятностей $p(x)$, $p(x,y_1)$ ограничены вместе со своими производными до второго порядка включительно; 3) ядерные функции $\Phi(u)$ являются положительными, симметричными и нормированными; 4) последовательности коэффициентов размытости $c_1(n)$, $c(n)$ ядерных функций таковы, что при $n\to\infty$ их значения стремятся к нулю. Тогда непараметрическая регрессия (3) обладает свойством асимптотической несмещённости.

Доказательство. Представим модель (3) в виде

$$\overline{\overline{y}}(x)=\frac{(nc_1c)^{-1}\sum_{i=1}^{n} y^i \Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\overline{y}_1-\overline{y}_1^i}{c}\right)}{(nc_1c)^{-1}\sum_{i=1}^{n}\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\overline{y}_1-\overline{y}_1^i}{c}\right)}=\frac{\overline{z}_1(x)}{\overline{z}_2(x)}. \qquad (4)$$

Проведем преобразования

$$M\frac{\overline{z}_1(x)}{\overline{z}_2(x)}= \\ =M\left[\frac{\overline{z}_1(x)}{M\,\overline{z}_2(x)}+\frac{\overline{z}_1(x)}{\overline{z}_2(x)\,M\,\overline{z}_2(x)}\left(M\,\overline{z}_2(x)-\overline{z}_2(x)\right)\right], \qquad (5)$$

где M – знак математического ожидания.

Ввиду ограниченности значений $\overline{\overline{y}}(x)=\frac{\overline{z}_1(x)}{\overline{z}_2(x)}$ свойства статистики (3) зависят от асимптотического поведения $M(\overline{z}_1(x))$, $M(\overline{z}_2(x))$. Вычислим

$$M\left(\overline{z}_2(x)\right)=(nc_1c)^{-1}\sum_{i=1}^{n}\int\ldots\int\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\times \\ \times\Phi\left(\sum_{v=1}^{k2}\alpha_v\frac{\overline{x}_{1v}-\overline{x}_{1v}^i}{c}\right)p\left(x_1^i,\overline{x}_{1v}^i,v=\overline{1,k2}\right)dx_1^i d\overline{x}_{1v}^i\ldots d\overline{x}_{1k2}^i.$$

Так как $x_1^i,\overline{x}_{1v}^i,v=\overline{1,k2}$ являются значениями одних и тех же случайных величин $t,t_v,v=\overline{1,k2}$ с плотностью вероятности $p\left(t,t_v,v=\overline{1,k2}\right)$, то

$$M\left(\overline{z}_2(x)\right)=(c_1c)^{-1}\int\ldots\int\Phi\left(\frac{x_1-t}{c_1}\right)\times \\ \times\Phi\left(\sum_{v=1}^{k2}\frac{\alpha_v}{c}(\overline{x}_{1v}-t_v)\right)p\left(t,t_v,v=\overline{1,k2}\right)dtdt_1\ldots dt_{k2}.$$

Проведём замену переменных $u=\frac{(x_1-t)}{c_1}$, $u_v=\frac{\alpha_v(\overline{x}_{1v}-t_v)}{c}$. После несложных преобразований получим

$$M\left(\overline{z}_2(x)\right)=\frac{c^{k2-1}}{\prod_{v=1}^{k2}\alpha_v}\int\ldots\int\Phi(u)\,\Phi\left(\sum_{v=1}^{k2}u_v\right)\times \\ \times p\left(x_1-c_1u,\overline{x}_{1v}-\frac{c}{\alpha_v}u_v,v=\overline{1,k2}\right)du\,du_1\ldots\,du_{k2}. \qquad (6)$$

Разложим функцию

$$p\left(x_1-c_1u,\overline{x}_{1v}-\frac{c}{\alpha_v}u_v,v=\overline{1,k2}\right)$$

в ряд Тейлора в точке $x=\left(x_1,x_{1v},v=\overline{1,k2}\right)$ и преобразуем (6) с учётом свойств:

$$\frac{1}{2^{k2-1}}\int_{-1}^{1}\ldots\int_{-1}^{1}\Phi\left(\sum_{v=1}^{k2}u_v\right)du_1\ldots du_{k2}=1,$$

$$\frac{1}{2^{k2-1}}\int_{-1}^{1}\ldots\int_{-1}^{1}u_t\Phi\left(\sum_{v=1}^{k2}u_v\right)du_1\ldots du_{k2}=0,\ t=\overline{1,k2},$$

$$\frac{1}{2^{k2-1}}\int_{-1}^{1}\ldots\int_{-1}^{1}u_t^2\,\Phi\left(\sum_{v=1}^{k2}u_v\right)du_1\ldots du_{k2}=\beta^2.$$

В результате при $n\to\infty$ имеем

$$M\left(\overline{z}_2(x)\right)\sim\frac{2^{k2-1}c^{k2-1}}{\prod_{v=1}^{k2}\alpha_v}\times \\ \times\left[p(x)+\frac{c_1^2}{2}p_{x_1}^{(2)}(x)+\frac{c^2\beta^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{\alpha_v^2}p_v^{(2)}(x)+0\left(c^4\right)\right], \qquad (7)$$

где $p_{x_1}^{(2)}(x)$, $p_v^{(2)}(x)$ – вторые производные плотности вероятности $p\left(x_1,\overline{x}_{1v},\nu=\overline{1,k2}\right)$ по переменным x_1, $\overline{x}_{1v},v=\overline{1,k2}$ соответственно.

Следуя приведённой технологии вычислений, найдём асимптотическое выражение для

$$M\left(\overline{\overline{z}}_1(x)\right)=\left(nc_1c\right)^{-1}\sum_{i=1}^{n}\int\ldots\int y^i\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\sum_{v=1}^{k2}\alpha_v\frac{\overline{x}_{1v}-\overline{x}_{1v}^i}{c}\right)\times$$

$$\times p\left(y^i,x_1^i,\overline{x}_{1v}^i,v=\overline{1,k2}\right)dy^i dx_1^i\,d\overline{x}_{11}^i\ldots d\overline{x}_{1k2}^i=$$

$$=\frac{c^{k2-1}}{\prod_{v=1}^{k2}\alpha_v}\int\ldots\int\varphi\left(x_1-c_1u,\overline{x}_{1v}-\frac{c}{\alpha_v}u_v,v=\overline{1,k2}\right)\times$$

$$\times\Phi(u)\,\Phi\left(\sum_{v=1}^{k2}u_v\right)p\left(x_1-c_1u,\overline{x}_{1v}-\frac{c}{\alpha_v}u_v,v=\overline{1,k2}\right)du\,du_1\ldots du_{k2}\sim$$

$$\sim\frac{2^{k2-1}\,c^{k2-1}}{\prod_{v=1}^{k2}\alpha_v}\left[\varphi(x)\,p(x)+\frac{c_1^2}{2}\left(\varphi(x)\,p(x)\right)_{x_1}^{(2)}+\right.$$

$$\left.+\frac{c^2\beta^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{\alpha_v^2}\left(\varphi(x)\,p(x)\right)_{x_{1v}}^{(2)}+0\left(c_1^2c^2,c_1^4c_v^4,v=\overline{1,k2}\right)\right]. \quad (8)$$

В выражении (8) $\left(\varphi(x)\,p(x)\right)_{x_1}^{(2)}$, $\left(\varphi(x)\,p(x)\right)_{x_{1v}}^{(2)}$ – вторые производные произведения двух функций по переменным x_1, $\overline{x}_{1v}, v=\overline{1,k2}$ соответственно.

Подставим выражения (7) и (8) в (5), получим

$$M\left(\overline{\overline{y}}(x)\right)\sim M\frac{\overline{\overline{z}}_1(x)}{\overline{\overline{z}}_2(x)}\sim$$

$$\sim\frac{\varphi(x)\,p(x)+\frac{c_1^2}{2}\left(\varphi(x)\,p(x)\right)_{x_1}^{(2)}+\frac{c^2\beta^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{\alpha_v^2}\left(\varphi(x)\,p(x)\right)_{x_{1v}}^{(2)}}{p(x)+\frac{c_1^2}{2}p_{x_1}^{(2)}(x)+\frac{c^2\beta^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{\alpha_v^2}p_v^{(2)}(x)}. \quad (9)$$

Из анализа выражения (9) следует, что при $c_1=c_1(n)\to 0$, $c=c(n)\to 0$ с ростом $n\to\infty$ изучаемая статистика (3) обладает свойством асимптотической несмещённости.

Замечания. При $k2=2$ полученные результаты могут быть использованы при исследовании свойств статистических моделей, основанных на методе группового учёта аргументов [5]. Идея метода заключается в построении последовательности моделей

$$\overline{y}_j=\overline{\varphi}_j(x_j,\overline{y}_{j-1}),i=\overline{1,m}. \quad (10)$$

Ранее не используемая в моделях $\overline{y}_t$, $t=\overline{1,j-1}$ компонента x_j вектора аргументов x обеспечивает в наборе с $\overline{y}_{j-1}$ минимальное расхождение значений $\overline{y}_j$ с экспериментальными данными. На каждом этапе процедуры (10) искомая зависимость оценивается в пространстве двух переменных $(x_j,\overline{y}_{j-1})$.

Анализ результатов вычислительных экспериментов. На основании данных вычислительных экспериментов сравнивалась эффективность статистики (2) и традиционной непараметрической регрессии

$$\tilde{y}(x)=\frac{\sum_{i=1}^{n}y^i\prod_{v=1}^{k}\Phi\left(\frac{x_v-x_v^i}{c_v}\right)}{\sum_{i=1}^{n}\prod_{v=1}^{k}\Phi\left(\frac{x_v-x_v^i}{c_v}\right)}. \quad (11)$$

В качестве искомой зависимости (1) использовался полином второй степени

$$\psi(x)=x_1^2+2x_2^2+x_1x_2+x_3+0,5x_4+2x_5, \quad (12)$$

каждый аргумент которого принимает значения из интервала $x_v\in[0;1]$, $v=\overline{1,5}$ с равномерным законом распределения. Частичные сведения о восстанавливаемой зависимости в соответствии с условиями теоремы определяются линейным полиномом

$$\overline{y}_1=x_3+0,5x_4+2x_5.$$

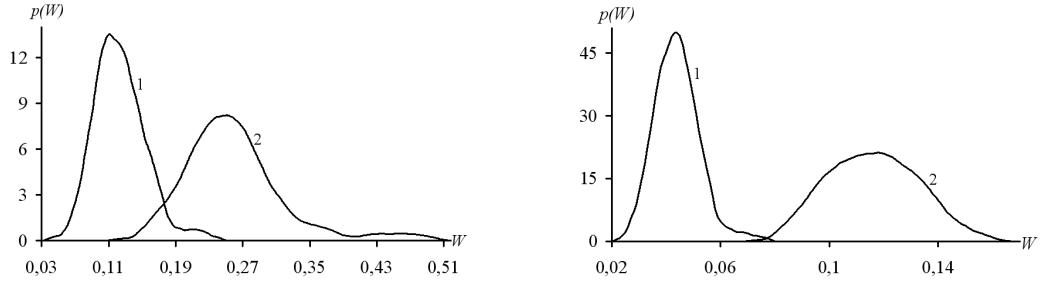
При формировании обучающей выборки $V=\left(x_v^i,v=\overline{1,5},y^i,i=\overline{1,n}\right)$ на значения функции (12) накладывалась аддитивная помеха

$$y^i=\psi\left(x^i\right)\left(1+2\left(\varepsilon^i-0,5\right)r\right), \quad (13)$$

где ε – случайная величина с равномерным законом в диапазоне $[0;1]$; r – параметр, определяющий уровень шума.

При синтезе непараметрических моделей (2), (11) использовалась ядерная функция Епанечникова, а их оптимизация по коэффициентам размытости осуществлялась в режиме «скользящего экзамена» из условия минимума среднеквадратического критерия. При этом полагалось, что значения коэффициентов размытости $c_v=c$, $v=\overline{1,5}$ для непараметрической регрессии и $\overline{c}_v=\overline{c}$, $v=1,2$ – для модели (2), так как интервалы изменения аргументов восстанавливаемой зависимости априори одинаковые. В качестве критерия эффективности моделей (2), (11) принимались среднеквадратические отклонения W_2, W_{11} их значений от функции (12), которые оценивались по контрольной выборке V_k объёма $n_k=10\,000$. При этом ситуации из выборки V_k, в которых исследуемые непараметрические модели не идентифицируют значения функции (12), не участвуют в формировании критериев их эффективности. Доля таких ситуаций не превышает значений 0,06 от объёма контрольной выборки.

Вычислительные эксперименты при фиксированных условиях исследования осуществлялись 60 раз. По полученным результатам восстанавливались плотности вероятности $p(W_2)$, $p(W_{11})$ оценок среднеквадратических отклонений W_2, W_{11} соответственно моделей (2), (11) (см. рисунок).



а *б*

Оценки плотностей вероятностей $p(W_2)$, $p(W_{11})$ среднеквадратических отклонений модифицированной непараметрической модели $\bar{\bar{y}}$ (кривая 1) и непараметрической регрессии $\tilde{y}$ (кривая 2). Условия эксперимента: объём обучающей выборки $n = 50$ (*а*), $n = 200$ (*б*); уровень шума в процедуре (13) $r = 0{,}1$

Анализ данных вычислительных экспериментов показывает, что статистические оценки законов распределения значений критериев эффективности W_2, W_{11} достоверно отличаются при различных объёмах обучающих выборок, причём интервалы изменения среднеквадратического критерия W_{11} непараметрической регрессии (11) характеризуются большими их значениями по сравнению с модифицированной регрессией (2). С ростом объёма обучающей выборки преимущество моделей (2) возрастает. Например, отношение $R = \bar{W}_{11}/\bar{W}_2$ средних значений W_2, W_{11} соответствует 2, 3 при $n = 50$ и 3 при $n = 200$.

Эффективность модифицированной непараметрической модели (2) объясняется возможностью снижения её размерности за счёт использования априорных сведений о наличии линейной взаимосвязи между переменными исследуемой зависимости. Данное заключение согласуется с результатами исследования гибридных моделей стохастических зависимостей [1].

Традиционная непараметрическая регрессия, основанная на оценке плотности вероятности типа Розенблатта–Парзена, обобщена при построении статистических моделей в условиях наличия частичных сведений о виде восстанавливаемых зависимостей. Предлагаемая модифицированная непараметрическая регрессия обладает свойством асимптотической несмещённости. Это позволяет аналитически обосновать возможность частичного сжатия пространства признаков на основе линейных преобразований без существенной потери полезной информации.

Перспективное направление дальнейших исследований состоит в развитии предлагаемого подхода на анализ свойств статистических моделей, основанных на методе группового учёта аргументов.

Библиографические ссылки

1. Лапко А. В., Лапко В. А. Гибридные модели стохастических зависимостей // Автометрия. 2002. № 5. С. 38–48.

2. Лапко В. А. Синтез и анализ гибридных моделей стохастических зависимостей в условиях наличия их частного описания // Автометрия. 2004. № 1. С. 51–59.

3. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. Vol. 33. P. 1065–1076.

4. Надарая Э. А. Непараметрические оценки кривой регрессии // Тр. ВЦ АН ГССР. 1965. Вып. 5. С. 56–68.

5. Ивахненко А. Г. Непараметрический комбинированный алгоритм МГУА на операторах поиска аналогов // Автоматика. 1990. № 5. С. 14–27.

A. V. Lapko, V. A. Lapko

NONPARAMETRIC REGRESSION IN THE CONDITIONS OF PARTIAL DATA ON A MODE OF RESTORED LEGITIMACIES

The nonparametric model of stochastic dependence which provides for registration of a priori data on a mode of restored legitimacies is offered. Its asymptotic properties are researched.

Keywords: nonparametric regression, stochastic dependences, a priori data, asymptotic properties.