УДК: 519.7

© 2020 г. **А.В. Лапко**^{1,2}, д-р техн. наук,

В.А. Лапко^{1,2}, д-р техн. наук

¹(Институт вычислительного моделирования СО РАН, Красноярск),

²(Сибирский государственный университет науки и технологий
имени академика М.Ф. Решетнева, Красноярск)

ПОСЛЕДОВАТЕЛЬНАЯ ПРОЦЕДУРА НЕПАРАМЕТРИЧЕСКОГО ОЦЕНИВАНИЯ СТОХАСТИЧЕСКОЙ ЗАВИСИМОСТИ

Предлагается методика построения непараметрической модели многомерной зависимости в условиях малых объемов исходных статистических данных. Основу методики составляет последовательная процедура формирования непараметрической модели в пространстве наборов аргументов восстанавливаемой зависимости. Исследуются асимптотические свойства разработанной модели и приводятся результаты вычислительных экспериментов.

Ключевые слова: непараметрическая регрессия, выборки малого объема, последовательная процедура формирования решений, асимптотические свойства.

DOI: 10.22250/isu.2020.66.95-103

Введение

Непараметрические модели, основанные на ядерных оценках плотности вероятности, используются при решении задач распознавания образов и восстановления многомерных зависимостей в условиях представительных статистических данных [1 – 6]. При решении прикладных задач часто встречаются выборки ограниченного объема наблюдений, когда в отношении «объем/размерность» (n / k) исходных мало. Подобные условия обусловлены нестационарностью объекта исследования, высокой стоимостью и сложностью получения необходимой информации.

Для восстановления стохастических зависимостей по данным малых выборок широко используются методы группового учета аргументов [7].

Основу подхода составляет последовательная процедура усложнения параметрической модели изучаемой зависимости путем селекции пар первичных и промежуточных ее аргументов. При этом каждый этап синтеза модели характеризуется отношением $n/k = n/2$.

Методика построения непараметрических моделей, основанная на декомпозиции исходной выборки по ее размерности с целью увеличения отношения n/k и принципах коллективного оценивания, предложена в работе [8]. В соответствии с предлагаемым подходом по результатам декомпозиции для каждого набора аргументов формируются частные решающие функции, которые обобщаются в линейном либо нелинейном коллективе. В первом случае весовые коэффициенты частных моделей определяются путем анализа их ошибок аппроксимации. Во втором случае обобщенная модель строится в пространстве значений оценок восстанавливаемой зависимости, которые рассчитываются по частным моделям.

Цель работы состоит в синтезе и исследовании последовательной непараметрической процедуры формирования обобщенной модели многомерной зависимости в условиях статистических данных малого объема. Каждый этап синтеза подобной модели основывается на корректировке ее промежуточного варианта в пространстве частного набора аргументов с помощью непараметрической регрессии.

Последовательная процедура формирования решений

Пусть при восстановлении многомерной стохастической зависимости $y = \varphi(x_1, \dots, x_k)$ имеется выборка $V = (x_1^i, \dots, x_k^i, y^i, i = \overline{1, n})$ малого объема n . Исходный набор компонент $x = (x_1, \dots, x_k)$ в соответствии со спецификой решаемой задачи допускает разбиение на группы $x(t), t = \overline{1, r}$.

Методика формирования последовательной процедуры принятия решений:

1. Построить модель искомой зависимости с использованием набора компонент $x(t)$ в виде непараметрической регрессии [9, 10]

$$\bar{\varphi}_t(x(t)) = \frac{\sum_{i=1}^n y^i \prod_{v \in I_t} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}{\sum_{i=1}^n \prod_{v \in I_t} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}, \quad (1)$$

где I_t – множество номеров компонент в группе $x(t)$.

Ядерные функции $\Phi(u_v)$ в статистике (1) удовлетворяют условиям [11]:

$$\Phi(u_v) = \Phi(-u_v), \quad 0 \leq \Phi(u_v) < \infty, \quad \int \Phi(u_v) du_v = 1, \quad \int u_v^2 \Phi(u_v) du_v = 1, \\ \int u_v^m \Phi(u_v) du_v < \infty, \quad 0 \leq m < \infty, \quad v = \overline{1, k}.$$

Здесь и далее бесконечные пределы интегрирования опускаются. Коэффициенты размытости $c_v = c_v(n)$, $v \in I_t$ ядерных функций убывают с ростом объема n выборки V .

Оптимизация непараметрической регрессии (1) по коэффициентам размытости ядерных функций c_v , $v \in I_t$ осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки среднеквадратической ошибки аппроксимации W_t непараметрической регрессии $\bar{\varphi}_t(x(t))$ при восстановлении искомой зависимости.

2. Повторить этап 1 для наборов компонент $x(t)$, $t = \overline{1, r}$ и выбрать набор $x^*(1)$, в пространстве которого непараметрическая регрессия $\bar{\varphi}_1(x^*(1))$ типа 1 имеет минимальную ошибку $W_1^* = \min_{t=\overline{1, r}} W_t$.

3. Сформировать массив

$$V_1 = (x^j(t), q_1^j = y^j - \bar{\varphi}_1(x^j(1)), j = \overline{1, n})$$

расхождений между наблюдениями y^j искомой зависимости и значениями непараметрической регрессии в условиях $x^j(t)$

$$\bar{\varphi}_1(x^{*j}(1)) = \frac{\sum_{i=1}^n y^i \prod_{v \in I_1^*} \Phi\left(\frac{x_v^j - x_v^i}{c_v}\right)}{\sum_{i=1}^n \prod_{v \in I_1^*} \Phi\left(\frac{x_v^j - x_v^i}{c_v}\right)},$$

где I_1^* – множество номеров компонент x_v из набора $x^*(1)$.

В массиве данных V_1 набор компонент $x(j)$ не совпадает с $x^*(1)$.

4. По данным V_1 построить непараметрическую модель корректирующей функции в пространстве набора компонент $x(t)$

$$\bar{q}_t(x(t)) = \frac{\sum_{j=1}^n q_1^j \prod_{v \in I_t} \Phi\left(\frac{x_v - x_v^j}{c_v}\right)}{\sum_{i=1}^n \prod_{v \in I_t} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}. \quad (2)$$

5. С учетом непараметрических моделей (1), (2) осуществить синтез фрагмента последовательной процедуры формирования решений

$$\bar{\varphi}_{1t}(x^*(1), x(t)) = \bar{\varphi}_1(x^*(1)) + \bar{q}_t(x(t)). \quad (3)$$

Обозначим через W_{1t} оценку средней квадратической ошибки аппроксимации статистикой (3) искомой зависимости $y = \varphi(x_1, \dots, x_k)$, вычисляемой в режиме «скользящего экзамена».

6. Повторить этапы (3) – (5) для набора компонент $x(t)$, $t = \overline{1, r}$, за исключением $x^*(1)$, и выбрать набор $x^*(2)$, в пространстве которого статистика (3) имеет минимальную ошибку аппроксимации $y = \varphi(x_1, \dots, x_k)$.

7. Результатом синтеза разработанного фрагмента последовательной процедуры принятия решений является статистика

$$\bar{\varphi}_{12}(x^*(1), x^*(2)) = \bar{\varphi}_1(x^*(1)) + \bar{q}_2(x^*(2)),$$

где $\bar{q}_2(x^*(2))$ определяется непараметрической регрессией (2) при $x(t) \neq x^*(1)$.

Далее, следуя технологии обработки полученных данных, формируется массив

$$V_{12} = (x^j(t), q_{12}^j = y^j - \bar{\varphi}_{12}(x^{*j}(1), x^{*j}(2)), j = \overline{1, n})$$

расхождений между наблюдениями y^j зависимости $y = \varphi(x_1, \dots, x_k)$ и значениями статистики $\bar{\varphi}_{12}(x^*(1), x^*(2))$ (3) в условиях $(x^{*j}(1), x^{*j}(2))$. При этом наборы компонент $x^*(1)$, $x^*(2)$ из анализа исключаются.

На основе массива данных V_{12} определяется непараметрическая оценка $\bar{q}_{12}(x(t))$, корректирующей функции типа (2) и новый фрагмент последовательной процедуры принятия решений

$$\bar{\varphi}_{12t}(x^*(1), x^*(2), x(t)) = \bar{\varphi}_{12}(x^*(1), x^*(2)) + \bar{q}_t(x(t)).$$

В соответствии с приведенной выше технологией синтеза последовательной процедуры формирования решений определяется набор компонент $x^*(3)$, обеспечивающий в пространстве $x^*(1)$, $x^*(2)$, $x(t)$ минимальную ошибку аппроксимации искомой зависимости $y = \varphi(x_1, \dots, x_k)$.

Тогда следующий фрагмент последовательной процедуры формирования решений запишется в виде

$$\bar{\varphi}_{123}(x^*(1), x^*(2), x^*(3)) = \bar{\varphi}_{12}(x^*(1), x^*(2)) + \bar{q}_3(x^*(3)).$$

Продолжая предлагаемый синтез последовательной непараметрической процедуры, получим следующую статистику в качестве модели $y = \varphi(x_1, \dots, x_k)$

$$\bar{\varphi}_{\overline{1, r}}(x^*(t), t = \overline{1, r}, x(r)) = \bar{\varphi}_{\overline{1, r-1}}(x^*(t), t = \overline{1, r-1}) + \bar{q}_r(x(r)).$$

На последнем этапе построения процедуры формирования решений набор признаков $x(r)$ является единственным, поэтому определяется без сравнения оценки ошибки аппроксимации.

Асимптотические свойства непараметрической модели типа (3) при $k = 2$

Будем полагать, что функции $\varphi_v(x_v)$, $p(x_v, y)$, $p(x_v)$, $v=1, 2$ в области определения $y = \varphi(x_1, x_2)$ ограничены и непрерывны со всеми своими производными до второго порядка включительно. Ядерные функции $\Phi(u)$ являются положительными, нормированными и симметричными, а также $\int u^m \Phi(u) du < \infty \forall m < \infty$ при конечных n .

В этих условиях рассмотрим асимптотические свойства непараметрической процедуры (3) при восстановлении зависимости $y = \varphi(x_1, x_2)$ вида

$$\bar{y} = \bar{\varphi}_1(x_1) + \bar{q}_1(x_2). \quad (4)$$

Составляющие статистики (4) в предположении, что плотности вероятностей $p(x_v)$, $v=1, 2$ известны, определяются выражениями

$$\begin{aligned} \bar{\varphi}_1(x_1) &= (n c_1 p(x_1))^{-1} \sum_{i=1}^n y^i \Phi\left(\frac{x_1 - x_1^i}{c_1}\right), \\ \bar{q}_1(x_2) &= (n c_2 p(x_2))^{-1} \sum_{j=1}^n (y^j - \bar{\varphi}_1(x_1^j)) \Phi\left(\frac{x_2 - x_2^j}{c_2}\right), \end{aligned}$$

которые формируются по выборке $V = (x_1^i, x_2^i, y^i, i = \overline{1, n})$.

Заметим, что (x_1^i, y^i) , (x_2^i, y^i) , $i = \overline{1, n}$, $x_1^j, x_2^j, j = \overline{1, n}$ являются наблюдениями одних и тех же случайных величин соответственно (t_1, y) , (t_2, y) , t_1', t_2' . Будем считать, что случайные величины t_1 и t_2 являются независимыми.

Тогда математическое ожидание статистики (4) запишется в виде

$$\begin{aligned} M(\bar{\varphi}(x_1, x_2)) &= \frac{1}{c_1 p(x_1)} \iint y p\left(\frac{y}{t_1}\right) dy \Phi\left(\frac{x_1 - t_1}{c_1}\right) p(t_1) dt_1 + \\ &+ \frac{1}{c_2 p(x_2)} \iint y p\left(\frac{y}{t_2}\right) dy \Phi\left(\frac{x_2 - t_2}{c_2}\right) p(t_2) dt_2 - \\ &- \left[\frac{n-1}{n p(x_1) p(x_2) c_1 c_2} \int \dots \int y p\left(\frac{y}{t_1}\right) \Phi\left(\frac{t_1' - t_1}{c_1}\right) \Phi\left(\frac{x_2 - t_2'}{c_2}\right) p(t_1) p(t_2') dy dt_1 dt_1' dt_2' + \right. \\ &\left. + \frac{1}{n p(x_1) p(x_2) c_1} \int y p(y) dy \int \Phi\left(\frac{x_2 - t_2'}{c_2}\right) p(t_2') dt_2' \right]. \end{aligned} \quad (5)$$

Вычислим первое слагаемое $M(\bar{\varphi}_1(x_1))$ в выражении (5). Условное математическое ожидание $\varphi_1(t_1) = \int y p\left(\frac{y}{t_1}\right) dy$ является оптимальной решаю-

щей функцией в смысле минимума среднеквадратического отклонения $\bar{\varphi}_1(x_1)$ от $\varphi_1(x_1)$. Проведем замену переменных $u_1 = \frac{x_1 - t_1}{c_1}$, $t_1 = x_1 - c_1 u_1$, $dt_1 = -c_1 du_1$, получим

$$M(\bar{\varphi}_1(x_1)) = \frac{1}{p(x_1)} \int \varphi_1(x_1 - c_1 u_1) \Phi(u_1) p(x_1 - c_1 u_1) du_1.$$

Разложим функции $\varphi_1(x_1 - c_1 u_1)$ и $p(x_1 - c_1 u_1)$ в ряды Тейлора в точке x_1 , имеем

$$M(\bar{\varphi}_1(x_1)) = \frac{1}{p(x_1)} \int \Phi(u_1) \left(\varphi_1(x_1) - c_1 u_1 \varphi_1^{(1)}(x_1) + \frac{c_1^2 u_1^2}{2} \varphi_1^{(2)}(x_1) + o(c_1^4) \right) \times \\ \times \left(p(x_1) - c_1 u_1 p^{(1)}(x_1) + \frac{c_1^2 u_1^2}{2} p^{(2)}(x_1) + o(c_1^4) \right) du_1.$$

Здесь $\varphi_1^{(1)}(x_1)$, $\varphi_1^{(2)}(x_1)$, $p^{(1)}(x_1)$, $p^{(2)}(x_1)$ – первые и вторые производные функции $\varphi_1(x_1)$, $p(x_1)$.

Учитывая, что $\int u_1^m \Phi(u_1) du_1 = 0$ при нечетных значениях m , и принимая $\int u_1^2 \Phi(u_1) du_1 = 1$, после несложных преобразований получим

$$M(\bar{\varphi}_1(x_1)) = \varphi_1(x_1) + c_1^2 A_1(x_1) + o(c_1^4), \quad (6)$$

$$\text{где } A_1(x_1) = \frac{p^{(2)}(x_1) \varphi_1(x_1)}{2 p(x_1)} + \frac{p^{(1)}(x_1) \varphi_1^{(1)}(x_1)}{p(x_1)} + \frac{\varphi_1^{(2)}(x_1)}{2}.$$

Выполняя аналогичные преобразования, нетрудно получить асимптотическое выражение для второго слагаемого (5)

$$M(\bar{\varphi}_2(x_2)) = \varphi_2(x_2) + c_2^2 A_2(x_2) + o(c_2^4), \quad (7)$$

где $\varphi_2(x_2)$ – модель зависимости $y = \varphi(x_1, x_2)$ в пространстве x_2 ;

$$A_2(x_2) = \frac{p^{(2)}(x_2) \varphi_2(x_2)}{2 p(x_2)} + \frac{p^{(1)}(x_2) \varphi_2^{(1)}(x_2)}{p(x_2)} + \frac{\varphi_2^{(2)}(x_2)}{2}.$$

Третье слагаемое в (5) является математическим ожиданием статистики

$$\bar{\bar{\varphi}}_1(x_2) = \frac{1}{n^2 p(x_1) p(x_2) c_1 c_2} \sum_{j=1}^n \sum_{i=1}^n y^i \Phi\left(\frac{x_1^j - x_1^i}{c_1}\right) \Phi\left(\frac{x_2 - x_2^j}{c_2}\right), \quad (8)$$

которая представляется в виде двух составляющих $\bar{\bar{\varphi}}_{11}(x_2)$ и $\bar{\bar{\varphi}}_{12}(x_2)$. Первая составляющая соответствует выражению (8) при $x_1^j = x_1^i$, а вторая для условий $x_1^i \neq x_1^j$, $i = \overline{1, n}$, $j = \overline{1, n}$.

Следуя предложенной выше технологии исследования асимптотических свойств непараметрических статистик, вычислим

$$\begin{aligned} M(\bar{\varphi}_{12}(x_2)) &\sim \frac{n-1}{n p(x_1) p(x_2)} \int \int \int \varphi_1(t'_1 - c_1 u_1) \Phi(u_1) p(t'_1 - c_1 u_1) p(t'_1) \times \\ &\quad \times \Phi(u_2) p(x_2 - c_2 u_2) du_2 dt'_1 du_1 = \\ &= \frac{A}{p(x_1)} + \frac{c_1^2 B}{2 p(x)} + \frac{c_2^2 A p^{(2)}(x_2)}{2 p(x_1) p(x_2)} + o\left(\frac{1}{n}\right) + o(c^4), \end{aligned} \quad (9)$$

где $A = \int \varphi_1(t'_1) p^2(t'_1) dt'_1$, $B = \int (\varphi_1(t'_1) p^{(2)}(t'_1) p(t'_1) + 2 \varphi_1^{(1)}(t'_1) p^{(1)}(t'_1) p(t'_1) + \varphi_1^2(t'_1) p^2(t'_1)) dt'_1$.

В процессе преобразований величины малости $1/n$, c^4 и выше не учитывались. Определим

$$\begin{aligned} M(\bar{\varphi}_{11}(x_2)) &= \frac{\Phi(0)}{n p(x_1) p(x_2) c_1 c_2} \int y p(y) dy \int \Phi\left(\frac{x_2 - t'_2}{c_2}\right) p(t'_2) dt'_2 \sim \\ &\sim \frac{M(y) p(x_2) \Phi(0)}{n c_1 p(x_1)} + o\left(\frac{c_2^2}{n c_1}\right), \end{aligned} \quad (10)$$

где $\Phi(0)$ – максимальное значение ядерной функции.

С учетом результатов (6),(7),(9),(10), математическое ожидание статистики $\bar{\varphi}(x_1, x_2)$ для условий $c_1 \rightarrow 0$, $c_2 \rightarrow 0$, а $n c_1 \rightarrow \infty$ при $n \rightarrow \infty$ запишется в виде

$$M(\bar{\varphi}(x_1, x_2)) \sim \varphi_1(x_1) + \varphi_2(x_2) - \frac{A}{p(x_1)}. \quad (11)$$

При равномерном законе распределения x_1 третье слагаемое выражения (11) определяется математическим ожиданием функции $\varphi_1(x_1)$.

Анализ результатов вычислительных экспериментов

Проведем анализ результата (11) и процедуры (4) при достаточно большом объеме n исходных статистических данных. В качестве искомой зависимости $y = \varphi(x_1, x_2)$ используется полином

$$y = \varphi(x_1, x_2) = x_1^2 + 0,5 x_2^2 + 5x_1 - 3x_2 + 10. \quad (12)$$

При формировании исходной выборки $V = (x_1^i, x_2^i, y^i, i = \overline{1, n})$ на значения функции накладывалась аддитивная помеха

$$y^i = \varphi(x_1^i, x_2^i) (1 + r(\varepsilon^i - 0,5)),$$

где ε^i – случайная величина с равномерным законом распределения на интервале $[0; 1]$. Параметр $r = 0,15$ определяет уровень помех.

Законы распределения x_1, x_2 являются гауссовскими $N(0; 1)$. Эффек-

тивность исследуемых моделей оценивалась среднеквадратической ошибкой аппроксимации на основе контрольной выборки объема 500. Значения непараметрических оценок функций $\varphi_1(x_1)$, $\varphi_2(x_2)$ вычислялись 10 раз и полученные оценки ошибок аппроксимации усреднялись.

Предлагаемая методика синтеза последовательной процедуры формирования решений (4) сравнивалась с результатом исследования асимптотических свойств статистики (11).

В этих условиях среднеквадратическая ошибка аппроксимации искомой зависимости статистикой (4) соответствует значению 0,0619, а асимптотическим выражением (11) – значению 0,0628. Полученные результаты подтверждают корректность выполненных аналитических преобразований при анализе последовательной процедуры непараметрического оценивания стохастических зависимостей.

Сформулированная задача исследования также рассматривалась при использовании параметрических моделей зависимости (12):

$$\begin{aligned}\varphi_1(x_1) &= 1,38x_1^2 + 4,87x_1 + 10,13, \\ \varphi_2(x_2) &= 0,31x_2^2 - 2,92x_2 + 11,26, \\ \varphi_1(x_2) &= 0,495x_2^2 - 2,94x_2 - 0,49.\end{aligned}\tag{13}$$

Их применение при синтезе последовательной процедуры формирования решений (4), (11) позволяет получить среднеквадратические ошибки аппроксимации, которые отличаются от показателей эффективности аппроксимации непараметрических моделей не более чем на 10%. Из анализа выражения (12) при использовании моделей (13) следует, что полученные положительные результаты могут объясняться отсутствием в искомой зависимости (12) составляющих в виде произведения x_1x_2 .

Заключение

Последовательная процедура непараметрического оценивания стохастических зависимостей позволяет их восстанавливать в условиях малых объемов исходных данных. Основу последовательной процедуры составляют декомпозиция аргументов восстанавливаемой зависимости на независимые их наборы, определение набора аргументов с минимальной ошибкой аппроксимации, формирование соответствующей частной непараметрической модели и корректирующей ее функции. Сумма частной модели и корректирующей ее функции определяет первый этап последовательной процедуры восстановления искомой зависимости. По аналогии, используя полученные результаты, формируется второй и последующие этапы синтеза не-

параметрической модели искомой зависимости. Двухмерная непараметрическая модель, синтез которой основан на предлагаемой методике, обладает свойством асимптотической несмещенности, что подтверждается результатами вычислительных экспериментов.

Дальнейшее развитие исследований в данном направлении связано с углубленным изучением асимптотических свойств результатов применения последовательной процедуры формирования непараметрической модели стохастической зависимости.

ЛИТЕРАТУРА

1. Лапко А.В., Лапко В.А. Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. – 2008. – Т. 44, № 3. – С. 65-74.
2. Лапко А.В., Лапко В.А., Егорочкин И.А. Непараметрические оценки смеси плотностей вероятности и их применение в задаче распознавания образов // Системы управления и информационные технологии. – 2009. – Т. 35, № 1. – С. 60-64.
3. Лапко А.В., Лапко В.А. Коллектив непараметрических решающих функций в двухальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2009. – Т. 37, № 3.1. – С. 156-160.
4. Лапко А.В., Лапко В.А. Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и ее свойства // Системы управления и информационные технологии. – 2010. – Т. 39, № 1.2. – С. 300-304.
5. Лапко А.В., Лапко В.А. Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. – 2012. – Т. 32, № 2. – С. 121-126.
6. Лапко В.А., Капустин А.Н. Синтез нелинейных непараметрических решающих правил в задачах распознавания образов // Автометрия. – 2006. – Т. 42, № 6. – С. 26-33.
7. Ивахненко А.Г. Непараметрический комбинированный алгоритм МГУА на операторах поиска аналогов // Автоматика. – 1990. – № 5. – С. 14-27.
8. Лапко А.В., Лапко В.А. Исследование непараметрической регрессии в условиях частичных сведений о виде восстанавливаемой зависимости // Информатика и системы управления. – 2011. – Т. 30, № 4. – С. 124-132.
9. Надарая Э.А. Непараметрические оценки кривой регрессии // Труды ВЦ АН ГССР. – 1965. – Вып. 5. – С. 56-68.
10. Хардле В. Прикладная непараметрическая регрессия. – М.: Мир, 1993. – 349 с.
11. Parzen E. On the estimation of probability density function and mode // Annals of Mathematical Statistics. – 1962. – Vol. 33. – P.1065-1076.

E-mail:

Лапко Александр Васильевич – lapko@ict.krasn.ru;

Лапко Василий Александрович – lapko@ict.krasn.ru.