УДК 519.7

РЕГРЕССИОННАЯ ОЦЕНКА МНОГОМЕРНОЙ ПЛОТНОСТИ ВЕРОЯТНОСТИ И ЕЁ СВОЙСТВА*

А. В. Лапко[1, 2], **В. А. Лапко**[1, 2]

[1] *Институт вычислительного моделирования СО РАН, 660036, г. Красноярск, Академгородок, 50, стр. 44*
[2] *Сибирский государственный аэрокосмический университет им. академика М. Ф. Решетнёва, 660014, г. Красноярск, просп. им. Газеты «Красноярский рабочий», 31*
E-mail: lapko@icm.krasn.ru

Предлагается методика синтеза и анализа регрессионной оценки многомерной плотности вероятности в условиях большого объёма исходных статистических данных. Исследуются её асимптотические свойства. На этой основе устанавливается зависимость между показателями эффективности предлагаемой оценки и параметрами процедуры декомпозиции исходных данных.

Ключевые слова: плотность вероятности, регрессионная оценка, большие выборки, асимптотические свойства, априорная информация.

Введение. Вычислительная эффективность непараметрических алгоритмов обработки информации во многом определяется объёмом статистических данных и снижается по мере его увеличения, что затрудняет построение систем принятия решений в условиях больших выборок.

Естественным выходом в такой ситуации является использование принципов декомпозиции исходных статистических данных по их объёму и технологии параллельных вычислений. С этих позиций предложена и исследована смесь непараметрических оценок плотностей вероятности для одномерных и многомерных случайных величин [1, 2]. Показано, что она имеет значительно меньшую дисперсию по сравнению с традиционной непараметрической оценкой плотности вероятности типа Розенблатта — Парзена [3]. При этом сокращение времени вычислений сопоставимо с количеством составляющих смеси непараметрических оценок плотностей вероятности.

Полученные результаты обобщены при оценивании решающей функции в задаче распознавания образов для условий больших выборок. Разработаны двухуровневые непараметрические системы для решения двуальтернативной [4] и многоальтернативной [5] задач классификации, установлены асимптотические свойства оценок их уравнений разделяющих поверхностей для одномерного и многомерного случаев [6].

Перспективное направление «обхода» проблем больших выборок связано с декомпозицией исходных статистических данных и последующим анализом количественных характеристик получаемых множеств случайных величин [7]. На этой основе построена регрессионная оценка плотности вероятности одномерной случайной величины [8].

Цель предлагаемой работы заключается в исследовании свойств регрессионной оценки многомерной плотности вероятности и определении их зависимости от особенностей процедуры декомпозиции статистических данных, что создаёт теоретическую основу разработки эффективных алгоритмов принятия решений в условиях больших выборок.

*Работа выполнена в рамках базовой части государственного задания Министерства образования и науки РФ высшим учебным заведениям на 2014–2016 гг. (СибГАУ № Б121/14).

Синтез оценки плотности вероятности. Пусть дана выборка $V = (x^j,\ j = \overline{1,n})$ из n независимых наблюдений многомерной случайной величины $x = (x_v,\ v = \overline{1,k})$ с неизвестной плотностью вероятности $p(x)$. Предполагается, что $p(x)$ разлагается в ряд Тейлора по всем своим аргументам в каждой точке x.

Для упрощения аналитических преобразований будем считать, что интервалы Δ_v изменения аргументов x_v равны, т. е. $\Delta_v = \Delta$, $v = \overline{1,k}$. Разобьём область изменения каждого аргумента на $\bar{N}$ непересекающихся интервалов длиной 2β таких, что $\Delta = 2\beta\bar{N}$. В этих условиях по исходным данным V сформируем множества X^i, $i = \overline{1,N}$, $N = \bar{N}^k$. В качестве характеристик X^i примем частоту P^i попадания случайной величины x в i-й многомерный интервал и его центр z^i. На основе полученной информации составим статистическую выборку $V_1 = (z^i,\ y^i = P^i/(2\beta)^k,\ i = \overline{1,N})$. Центры z введённых многомерных интервалов имеют равномерный закон распределения $p_2(z) = (N(2\beta)^k)^{-1}$. Объём N полученной выборки может быть значительно меньше объёма n исходных статистических данных.

В качестве приближения по эмпирическим данным искомой плотности $p(x)$ примем статистику

$$\bar{p}(x) = \frac{1}{Nc^k p_2(z)} \sum_{i=1}^{N} y^i \prod_{v=1}^{k} \Phi\Big(\frac{x_v - z_v^i}{c}\Big), \tag{1}$$

которая является непараметрической оценкой условного математического ожидания $\varphi(x) = \int yp(y/x)dy$. Здесь и далее бесконечные пределы интегрирования опускаются.

В регрессионной оценке плотности вероятности (1) ядерные функции $\Phi(u_v)$ удовлетворяют условиям

$$\Phi(u_v) = \Phi(-u_v); \quad 0 \le \Phi(u_v) < \infty; \quad \int \Phi(u_v)du_v = 1; \quad \int u_v^2\Phi(u_v)du_v = 1, \tag{2}$$

а их коэффициенты размытости $c = c(N)$ убывают с ростом N.

Нетрудно убедиться, что регрессионная оценка плотности $\bar{p}(x)$ является нормированной функцией, т. е. удовлетворяет основному свойству плотности вероятности.

Асимптотические свойства регрессионной оценки многомерной плотности вероятности. При анализе асимптотических свойств статистики (1) будем использовать технологию преобразований, предложенную в работе [9] и развитую в [10–15].

1. По определению имеем

$$M(\bar{p}(x)) = \frac{1}{Nc^k p_2(z)} \sum_{i=1}^{N} \int \ldots \int y^i \prod_{v=1}^{k} \Phi\Big(\frac{x_v - z_v^i}{c}\Big) p(y^i, z_1^i, \ldots, z_k^i) dy^i dz_1^i \ldots dz_k^i =$$

$$= \frac{1}{c^k p_2(z)} \int \ldots \int y \prod_{v=1}^{k} \Phi\Big(\frac{x_v - t_v}{c}\Big) p(y, t_1, \ldots, t_k) dy dt_1 \ldots dt_k =$$

$$= \frac{1}{c^k} \int \ldots \int \varphi(t_1, \ldots, t_k) \prod_{v=1}^{k} \Phi\Big(\frac{x_v - t_v}{c}\Big) dt_1 \ldots dt_k, \tag{3}$$

где M — знак математического ожидания. При выполнении данных преобразований учитывается, что элементы статистической выборки V_1 являются значениями одних и тех же случайных величин $(t_1, \ldots, t_k, y)$ с плотностью вероятности $p(y, t_1, \ldots, t_k)$.

Проведём в выражении (3) замену переменных $(x_v - t_v)c^{-1} = u_v$ и разложим функцию $\varphi(x_v - cu_v,\ v = \overline{1,k})$ в ряд Тейлора в точке x. Тогда с учётом свойств (2) ядерных функций при достаточно больших значениях N получим асимптотическое выражение смещения регрессионной оценки многомерной плотности вероятности от оптимальной решающей функции $\varphi(x)$:

$$M(\bar{p}(x) - \varphi(x)) \sim \frac{c^2}{2} \sum_{v=1}^{k} \varphi_v^{(2)}(x), \tag{4}$$

где $\varphi_v^{(2)}(x)$ — вторая производная функции $\varphi(x)$ по аргументу x_v. Отсюда при условии $c \to 0$, когда $N \to \infty$, следует свойство асимптотической несмещённости регрессионной оценки многомерной плотности вероятности (1).

2. Исследуем асимптотические свойства среднеквадратического отклонения

$$M(\bar{p}(x) - \varphi(x))^2 = M(\bar{p}^2(x)) - 2\varphi(x)M(\bar{p}(x)) + \varphi^2(x). \tag{5}$$

Следуя ранее использованной технологии исследований, проведём преобразования

$$M(\bar{p}^2(x)) = \frac{1}{N^2 c^{2k} p_2^2(z)} \Big[\sum_{i=1}^{N} M\Big((y^i)^2 \prod_{v=1}^{k} \Phi^2\Big(\frac{x_v - z_v^i}{c}\Big)\Big) +$$

$$+ \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} M\Big(y^i \prod_{v=1}^{k} \Phi\Big(\frac{x_v - z_v^i}{c}\Big) y^j \prod_{v=1}^{k} \Phi\Big(\frac{x_v - z_v^j}{c}\Big)\Big)\Big] =$$

$$= \frac{1}{N c^{2k} p_2(z)} \int \dots \int \varphi^2(t_1, \dots, t_k) \prod_{v=1}^{k} \Phi^2\Big(\frac{x_v - t_v}{c}\Big) dt_1 \dots dt_k +$$

$$+ \frac{N-1}{N c^{2k}} \Big(\int \dots \int \varphi(t_1, \dots, t_k) \prod_{v=1}^{k} \Phi\Big(\frac{x_v - t_v}{c}\Big) dt_1 \dots dt_k \Big)^2.$$

Пренебрегая величинами малости $0(1/N)$, $0(1/(Nc^{k-2}))$, $0(c^6)$ и выше, найдём асимптотическое выражение

$$M(\bar{\varphi}^2(x)) \sim \varphi^2(x) + \frac{1}{N c^k p_2(z)} \varphi^2(x) \prod_{v=1}^{k} \int \Phi^2(u_v) du_v +$$

$$+ \frac{c^4}{4} \Big(\sum_{v=1}^{k} \varphi_v^{(2)}(x) \Big)^2 + c^2 \varphi(x) \sum_{v=1}^{k} \varphi_v^{(2)}(x). \tag{6}$$

Подставим выражения (4), (7) в (6) и при достаточно больших N получим

$$M(\bar{p}(x) - \varphi(x))^2 \sim \frac{\varphi^2(x)}{N c^k p_2(z)} \prod_{v=1}^{k} \int \Phi^2(u_v) du_v + \frac{c^4}{4} \Big(\sum_{v=1}^{k} \varphi_v^{(2)}(x) \Big)^2. \tag{7}$$

Отсюда, принимая во внимание соотношения (4) и (7), из условий $c \to 0$, $Nc^k \to \infty$ при $N \to \infty$ следуют свойства сходимости в среднеквадратическом и состоятельности регрессионной оценки многомерной плотности вероятности (1).

При $k = 1$ полученный результат (7) совпадает с утверждением работы [8], что подтверждает корректность выполненных преобразований.

Анализ аппроксимационных свойств статистики $\bar{p}(x)$. Основываясь на результатах аналитических исследований, установим количественную зависимость аппроксимационных свойств $\bar{p}(x)$ от параметров регрессионной оценки многомерной плотности вероятности и особенностей статистических данных V_1. Будем считать, что ядерные функции $\Phi(u_v)$, $v = \overline{1,k}$, одинаковы и равны $\Phi(u)$.

С учётом равномерного закона распределения $p_2(z)$ найдём минимальное значение выражения

$$M \int (\bar{p}(x) - \varphi(x))^2 dx \sim \frac{(\|\Phi(u)\|^2 2\beta)^k \|\varphi(x)\|^2}{c^k} + \frac{c^4}{4} B, \tag{8}$$

полученного путём интегрирования результата (7). Здесь $\|\Phi(u)\|^2 = \int \Phi^2(u)du$; $\|\varphi(x)\|^2 = \int \ldots \int \varphi^2(x_1, \ldots, x_k)dx_1 \ldots dx_k$; $B = \int \ldots \int \Big(\sum\limits_{v=1}^{k} \varphi_v^{(2)}(x_1, \ldots, x_k) \Big)^2 dx_1 \ldots dx_k$.

Для этого из условия минимума (8) по коэффициенту размытости c вычислим его оптимальное значение

$$\bar{c} = (k\|\varphi(x)\|^2 (2\beta \|\Phi(u)\|^2)^k / B)^{1/(k+4)}. \tag{9}$$

Подставив $\bar{c}$ (9) в выражение (8), запишем его минимальное значение

$$W_2 = \Big[(\|\varphi(x)\|^2 (2\beta \|\Phi(u)\|^2)^k)^4 B^k \Big]^{1/(k+4)} \frac{4+k}{4k^{k/(k+4)}}. \tag{10}$$

Пусть восстанавливаемая многомерная плотность вероятности имеет вид

$$p(x) = 1/(2\pi)^{k/2} \prod_{v=1}^{k} \exp(-x_v^2/2),$$

оптимальная модель которой в смысле минимума среднеквадратического отклонения задаётся выражением $\varphi(x)$. Поэтому будем считать, что $\varphi(x) \approx p(x)$. Тогда $\|\varphi(x)\|^2 \approx \approx 1/(2\sqrt{\pi})^k$, а значение B определяется выражением $B \approx k(2+k)/(4(2\sqrt{\pi})^k)$ [9].

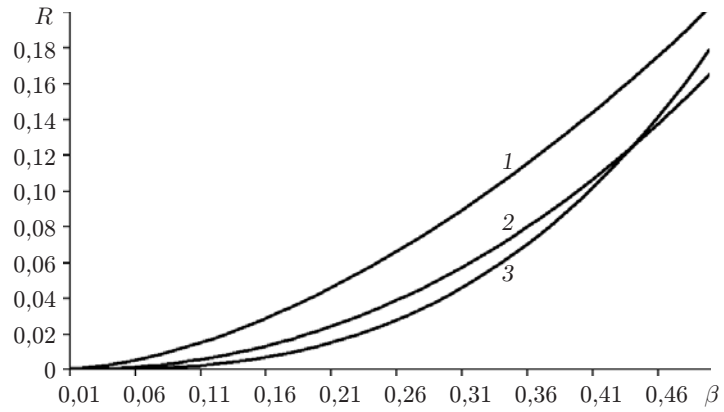
В качестве ядерной функции $\Phi(u)$ используем оптимальное ядро Епанечникова [9]

$$\Phi(u) = \begin{cases} \dfrac{3}{4\sqrt{5}} - \dfrac{3u^2}{20\sqrt{5}}, & \text{если } |u| < \sqrt{5}, \\ 0, & \text{если } |u| \geq \sqrt{5}, \end{cases}$$

для которого $\|\Phi(u)\|^2 = 3/(5\sqrt{5})$.

Тогда выражение (10) запишется в виде

$$W_2 = \frac{4+k}{4} \Big[\Big(\frac{3\beta}{5\sqrt{5\pi}} \Big)^4 \Big(\frac{2+k}{4(2\sqrt{\pi})^k} \Big)^k \Big]^{k/(k+4)}.$$



Зависимость относительной ошибки аппроксимации R оптимальной решающей функции $\varphi(x)$ регрессионной оценкой плотности $\bar{p}(x)$ от параметра дискретизации β области изменения многомерной случайной величины и её размерности k. Кривые *1–3* соответствуют значениям $k = 3; 5; 10$

В данных условиях исследуем относительную ошибку аппроксимации $R = W_2/\|\varphi(x)\|^2$.

С уменьшением значений параметра β процедуры дискретизации области изменения многомерной случайной величины наблюдается снижение относительной ошибки аппроксимации R статистики $\bar{p}(x)$ (см. рисунок). Данный факт объясняется увеличением объёма N выборки V_1, используемой при оценивании кривой регрессии $\varphi(x)$ с помощью статистики (1), и согласуется с условиями асимптотической сходимости $\bar{p}(x)$. С ростом β темп увеличения R особенно значителен в интервале больших значений β и размерности k многомерной случайной величины. При значениях $\beta < 0{,}2$ темп снижения R слабо зависит от изменения β и размерности k случайной величины.

Из анализа представленных зависимостей следует, что при фиксированных R с ростом размерности k случайной величины повышаются значения параметра дискретизации β. Например, при $R = 0{,}02$ значениям $k = 3; 5; 10$ соответствуют значения $\beta = 0{,}13; 0{,}2; 0{,}24$.

На основе полученных результатов появляется возможность найти аналитическую зависимость параметра дискретизации β от объёма n исходных статистических данных V, что актуально не только при синтезе регрессионной оценки плотности вероятности $\bar{p}(x)$ (1), но и при проверке гипотезы о тождественности законов распределения многомерных случайных величин с использованием критерия Пирсона [16].

Для этого необходимо исследовать асимптотические свойства регрессионной оценки плотности вероятности (1), представленной в виде

$$\bar{p}(x) = \frac{1}{c^k}\sum_{i=1}^{N} P^i \prod_{v=1}^{k} \Phi\Big(\frac{x_v - z_v^i}{c}\Big) = \frac{1}{nc^k}\sum_{i=1}^{N}\sum_{j=1}^{n}\prod_{v=1}^{k} h\Big(\frac{x_v^j - z_v^i}{\beta}\Big)\prod_{v=1}^{k}\Phi\Big(\frac{x_v - z_v^i}{c}\Big), \qquad (11)$$

где индикаторная функция

$$h\Big(\frac{x_v^j - z_v^i}{\beta}\Big) = \begin{cases} 1, & \text{если } |x_v^j - z_v^i| \le \beta, \\ 0, & \text{если } |x_v^j - z_v^i| > \beta, \end{cases}$$

определяет принадлежность элементов выборки $V = (x^j,\ j = \overline{1,n})$ многомерным интервалам $(z_v^i \pm \beta,\ v = \overline{1,k})$, $i = \overline{1,N}$. В отличие от (1) в её модификации (11) в явном виде присутствуют параметры процедуры дискретизации β, N и объёма n исходных статистических

данных. Поэтому в результате анализа статистики (11) с использованием предложенной технологии преобразований может быть получено соответствующее $\bar{p}(x)$ асимптотическое выражение среднеквадратического отклонения $W_2(c, N)$, зависящее от коэффициента размытости ядерных функций регрессионной оценки плотности (11) и количества интервалов дискретизации области изменения случайной величины. Подставляя в $W_2(c, N)$ оптимальное значение $\bar{c}$ (9), получим выражение $W_2(N)$, минимизация которого по параметру N даёт аналитическую зависимость количества N интервалов дискретизации от объёма n исходных статистических данных.

Заключение. Регрессионная оценка плотности вероятности является эффективным средством обработки статистических данных большого объёма. Её синтез осуществляется путём декомпозиции исходной информации и анализа на основе кривой регрессии количественных характеристик получаемых множеств случайных величин. Предлагаемая статистика обладает свойством асимптотической несмещённости и состоятельности относительно оптимального решающего правила.

Установлена зависимость относительной ошибки аппроксимации оптимальной решающей функции регрессионной оценкой плотности вероятности от параметра дискретизации области изменения многомерной случайной величины и её размерности.

Полученные результаты открывают возможность определения аналитической зависимости количества многомерных интервалов дискретизации от объёма исходных статистических данных, что имеет важное значение в задачах проверки гипотез о распределениях многомерных случайных величин с использованием критерия Пирсона.

СПИСОК ЛИТЕРАТУРЫ

1. **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. **35**, № 1. С. 60–64.
2. **Лапко А. В., Лапко В. А.** Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. 2011. **43**, № 1. С. 12–15.
3. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Stat. 1962. **33**, N 3. P. 1065–1076.
4. **Лапко А. В., Лапко В. А.** Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. 2009. **37**, № 3.1. С. 156–160.
5. **Лапко А. В., Лапко В. А.** Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. 2010. **46**, № 1. С. 70–78.
6. **Лапко А. В., Лапко В. А.** Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. 2011. **47**, № 4. С. 76–82.
7. **Лапко А. В., Лапко В. А.** Непараметрические методики анализа множеств случайных величин // Автометрия. 2003. **39**, № 1. С. 54–61.
8. **Лапко А. В., Лапко В. А.** Регрессионная оценка плотности вероятности и её свойства // Системы управления и информационные технологии. 2012. **49**, № 3.1. С. 152–156.
9. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, № 1. С. 156–161.
10. **Лапко А. В., Лапко В. А.** Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. 2012. **32**, № 2. С. 121–126.

11. **Лапко А. В., Лапко В. А.** Анализ дисперсии среднеквадратической ошибки аппроксимации непараметрической оценки плотности вероятности ядерного типа // Информатика и системы управления. 2012. **33**, № 3. С. 132–139.
12. **Лапко А. В., Лапко В. А.** Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов // Автометрия. 2010. **46**, № 3. С. 48–53.
13. **Лапко А. В., Лапко В. А.** Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и её свойства // Системы управления и информационные технологии. 2010. **39**, № 1.2. С. 300–304.
14. **Лапко А. В., Лапко В. А.** Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности в алгоритме распознавания образов, соответствующего критерию максимума апостериорной вероятности // Системы управления и информационные технологии. 2010. **42**, № 4. С. 58–61.
15. **Лапко А. В., Лапко В. А.** Анализ асимптотических свойств многомерной непараметрической регрессии // Вестн. СибГАУ. 2012. **42**, № 2. С. 41–44.
16. **Лапко А. В., Лапко В. А.** Сравнение непараметрических критериев проверки гипотез о распределениях случайных величин // Вестн. СибГАУ. 2011. **37**, № 4. С. 48–52.

Поступила в редакцию 2 апреля 2013 г.