519.24

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ГРАНИЦ ДЛЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ НА ОСНОВЕ ЕЕ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ

А. В. ЛАПКО, В. А. ЛАПКО***

** Институт вычислительного моделирования СО РАН, Красноярск, Россия*
*** Сибирский государственный аэрокосмический университет им. Акад. М. Ф. Решетнева, Красноярск, Россия, e-mail: lapko@icm.krasn.ru*

Предложена методика построения доверительных границ для плотности вероятности и исследована ее эффективность в зависимости от вероятностных характеристик процедуры декомпозиции области изменения случайной величины.

Ключевые слова: *плотность вероятности, доверительные границы, регрессионная оценка, декомпозиция исходных данных, непараметрическая статистика.*

The method of construction of confidential boundaries for probability density is suggested and its efficiency related to probability characteristics of procedure of decomposition of the range of random quantity variation is studied.

Key words: *probability density, confidential boundaries, regression estimation, decomposition of prior data, nonparametric statistics.*

Пусть имеется выборка $V = \left(x^i,\ i = \overline{1,\ n}\right)$ из n независимых значений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$. Для условий выборок большого объема n разработан ряд модификаций непараметрических оценок смеси плотностей вероятности [1—4]. Перспективным направлением решения проблем больших выборок является использование процедуры сжатия исходных статистических данных [5, 6].

Разобьем область определения $p(x)$ на N непересекающихся интервалов длиной 2β и сформируем множества случайных величин X^j, $j = \overline{1, N}$. В качестве характеристик X^j примем частоту $\overline{P}^j$ попадания случайной величины x в j-й интервал и его центр z^j. На основе

полученной информации определим массив данных $V_1 = \left(z^j,\ y^j = \overline{P}^j\ /\ (2\beta),\ j = \overline{1,\ N}\right)$, составленный из центров z^j введенных интервалов и соответствующих им значений оценок y^j плотности вероятности. Объем N полученных данных V_1 может быть значительно меньше объема n исходных статистических данных V.

В качестве приближения по эмпирическим данным V_1 искомой плотности вероятности $p(x)$ примем статистику [6]:

$$\tilde{p}(x) = c^{-1} \sum_{j=1}^{N} \overline{P}^j\, \Phi\left(\frac{x - z^j}{c}\right), \qquad (1)$$

в которой ядерные функции $\Phi(u)$ удовлетворяют условиям [7, 8]:

$$\Phi(u) = \Phi(-u);\quad 0 \le \Phi(u) < \infty;$$

$$\int_{-\infty}^{+\infty} \Phi(u)\, du = 1,\quad \int_{-\infty}^{+\infty} u^2 \Phi(u)\, du = 1,$$

а выбор коэффициентов размытости c ядерных функций в (1) характеризует качество приближения $p(x)$.

В данной работе предложен и исследован алгоритмический подход построения доверительных границ для $p(x)$ на основе ее непараметрической оценки (1).

Методика построения доверительных границ для плотности вероятности. Известно, что верхняя $P^j_{\text{в}}$ и нижняя $P^j_{\text{н}}$ границы интервальной оценки вероятности P^j события $x \in X^j$ с коэффициентом доверия γ определяются выражениями [9]:

$$P^j_{\text{в}} = \overline{P}^j + \frac{u_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{P}^j \left(1 - \overline{P}^j\right)}; \qquad (2)$$

$$P^j_{\text{н}} = \overline{P}^j - \frac{u_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{P}^j \left(1 - \overline{P}^j\right)}, \qquad (3)$$

где $u_{1-\alpha/2}$ — квантиль уровня $1-\alpha/2$ стандартного нормального распределения, его значения находят по таблицам квантилей нормального распределения при $\alpha = 1 - \gamma$.

Организуем вычислительный эксперимент и определим по $\overline{P}^{\,j}$, $j = \overline{1, N}$ в соответствии с (2), (3), значения $P_{\text{в}}^{j}, P_{\text{н}}^{j}$, $j = \overline{1, N}$, по которым осуществим синтез верхней и нижней границ $p(x)$:

$$\tilde{p}_{\text{в}}(x) = c_{\text{в}}^{-1} \sum_{j=1}^{N} \overline{P}_{\text{в}}^{\,j} \Phi\left(\frac{x - z^j}{c_{\text{в}}}\right); \tag{4}$$

$$\tilde{p}_{\text{н}}(x) = c_{\text{н}}^{-1} \sum_{j=1}^{N} \overline{P}_{\text{н}}^{\,j} \Phi\left(\frac{x - z^j}{c_{\text{н}}}\right). \tag{5}$$

Для выбора оптимальных коэффициентов размытости ядерных функций в (4), (5) воспользуемся методом «скользящего экзамена». В этом случае оптимальное значение коэффициента размытости $c_{\text{в}}$ ядерных функций, например, для верхней границы (4) будем выбирать из условия минимизации выражения

$$\sum_{i=1}^{N} \left(\frac{P_{\text{в}}^{i}}{2\beta} - \tilde{\tilde{p}}_{\text{в}}\left(z^i\right) \right)^2,$$

где $P_{\text{в}}^{i}/(2\beta)$ — верхняя доверительная граница $p(x)$ в пределах i-го интервала дискретизации случайной величины x, а

$$\tilde{\tilde{p}}_{\text{в}}\left(z^i\right) = \frac{1}{c_{\text{в}}} \sum_{\substack{j=1 \\ j \neq i}}^{N} P_{\text{в}}^{\,i} \Phi\left(\frac{z^i - z^j}{c_{\text{в}}}\right)$$

представляет ее оценку на этом интервале.

Анализ эффективности методики построения доверительных границ.

Исследуем влияние объема n статистических данных $V = \left(x^i,\ i = \overline{1, n}\right)$ и параметров процедуры их декомпозиции при построении доверительных границ для плотности вероятности с нормальным законом распределения

$$p(x) = \left(\sqrt{2\pi}\right)^{-1} \exp\left(-x^2 / 2\right). \tag{6}$$

Для выбора количества интервалов дискретизации области изменения значений случайной величины будем использовать формулы Хайнкольда—Гаеде и Брукса—Каррузера, соответственно:

$$N = \sqrt{n}; \tag{7}$$

$$N = 5\lg n. \tag{8}$$

Свойственные им зависимости N от n значительно отличаются при больших объемах n исходных статистических данных.

Синтез непараметрической оценки плотности вероятности (1) осуществлялся на основе ядерных функций В. А. Епанечникова [8]:

$$\Phi(u) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3u^2}{20\sqrt{5}} & \forall\ |u| < \sqrt{5}; \\ 0 & \forall\ |u| \geq \sqrt{5}. \end{cases}$$

При формировании границ $P_{\text{в}}^{j}$, $P_{\text{н}}^{j}$, $j = \overline{1, N}$ доверительные интервалы для вероятностей P^j, $j = \overline{1, N}$ определяли с учетом $\gamma = 0{,}9$; $0{,}95$.

При одних и тех же объемах n статистических данных V в соответствии с (4), (5) доверительные границы для $p(x)$ находили многократно ($m = 100$). В каждом вычислительном эксперименте устанавливался факт принадлежности $p(x)$ области $\Omega_{\text{н-в}}$, ограниченной границами (4), (5). Плотность вероятности $p(x)$ находится в пределах области $\Omega_{\text{н-в}}$, если она не имеет ни одного пересечения с границами $\tilde{p}_{\text{в}}(x)$, $\tilde{p}_{\text{н}}(x)$. По полученной информации оценивали вероятность $\bar{\gamma}_p$ попадания $p(x)$ в область $\Omega_{\text{н-в}}$. Значения $\bar{\gamma}_p$ можно определить как оценку коэффициента доверия при построении доверительных границ для $p(x)$.

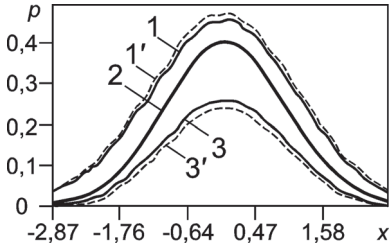
Установлено, что при малых $n \leq 150$ значения оценок $\bar{\gamma}_p$ сопоставимы в случае использования (7), (8) и количество N интервалов

дискретизации, вычисленных по этим формулам, отличается незначительно. С увеличением объема исходных статистических данных V применение (7) позволяет построить доверительные границы для $p(x)$ с более высоким $\bar{\gamma}_p$ (см. таблицу). Отмеченная закономерность особенно характерна для больших объемов n. Она объясняется более высокой «чувствительностью» формулы (7) к увеличению n по сравнению с (8). Увеличение N способствует повышению аппроксимационных свойств непараметрической оценки (1), которая является основой построения доверительных границ (4), (5). Однако при этом ухудшаются условия оценивания вероятностей попадания случайной величины в интервалы дискретизации, что снижает качество восстановления $p(x)$. По-видимому, существует определенное соотношение между объемом n исходных статистических данных и количеством N интервалов дискретизации, при котором достигаются высокие аппроксимационные свойства непараметрической оценки (1). При дальнейшем увеличении N следует ожидать снижение точности оценки $p(x)$. Полученные выводы оказываются достаточно общими и могут быть распространены на методы дискретизации интервала изменения случайной величины, для которых характер зависимости N от n является близким к (7), (8).

Оценки $\bar{\gamma}_p$ при построении доверительных границ для плотности вероятности $p(x)$ по (6)

γ; формула выбора N	Коэффициенты доверия $\bar{\gamma}_p$ при объеме n статистических данных									
	50	100	150	200	250	300	350	400	450	500
0,90; (7)	0,75	0,68	0,71	0,76	0,81	0,83	0,77	0,84	0,85	0,79
0,90; (8)	0,81	0,68	0,66	0,50	0,28	0,20	0,15	0,12	0,07	0,03
0,95; (7)	0,92	0,91	0,89	0,91	0,95	0,97	0,99	0,99	0,98	0,97
0,95; (8)	0,98	0,91	0,85	0,74	0,63	0,48	0,44	0,29	0,19	0,10

Соотношение $\bar{\gamma}_p \le \gamma$ справедливо при построении доверительных границ для $p(x)$ в случае $\gamma = 0{,}9$. С ростом γ увеличиваются размеры области $\Omega_{\text{н-в}}$, определяемые доверительными границами (4), (5). Поэтому при $\gamma = 0{,}95$ и использовании (7) в условиях $n > 200$ оценки



Доверительные границы для плотности вероятности $p(x)$ (кривая *2*) при $\gamma = 0{,}9$ (кривые *1*, *3*) и $\gamma = 0{,}95$ (кривые *1′*, *3′*). Метод дискретизации области изменения случайной величины определяется формулой (7); $n = 200$

$\bar{\gamma}_p$ сопоставимы со значениями γ. Примеры доверительных границ, соответствующих различным значениям γ при конкретных условиях вычислительного эксперимента, приведены на рисунке.

С увеличением объема n статистических данных доверительные границы для $p(x)$ сужаются, что согласуется с результатами исследования асимптотических свойств непараметрической оценки $p(x)$ [6].

Выводы. Структура рассматриваемой непараметрической оценки плотности вероятности позволяет решить проблему ее доверительного оценивания. Предлагаемый подход предполагает разбиение области значений случайной величины x на непересекающиеся интервалы Δ_j, $j = \overline{1, N}$ и последующее доверительное оценивание вероятностей принадлежностей x данным интервалам по исходной статистической информации. На этой основе можно осуществить синтез доверительных границ плотности вероятности. Размеры области, определяемые доверительными границами, зависят от количества N интервалов дискретизации значений случайной величины, объема n исходных статистических данных и заданного коэффициента доверия γ. При больших n применение формулы Хайнкольда—Гаеде для дискретизации интервала изменения случайной величины является предпочтительным. При относительно малых n результаты использования формул Хайнкольда—Гаеде и Брукса—Каррузера сопоставимы.

Предложенный подход открывает возможность использовать полученные результаты при построении доверительных границ многомерной плотности вероятности.

ЛИТЕРАТУРА

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. Т. 35. № 1. С. 60—64.

2. **Лапко А. В., Лапко В. А.** Анализ свойств смеси непараметрических оценок плотности вероятности многомерной случайной величины // Вестник СибГАУ. 2010. Т. 28. № 2. С. 32—35.

3. **Лапко А. В., Лапко В. А.** Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. 2011. Т. 47. № 4. С. 76—82.

4. **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.

5. **Лапко А. В., Лапко В. А.** Непараметрические методики анализа множеств случайных величин // Автометрия. 2003. Т. 39. № 1. С. 54—61.

6. **Лапко А. В., Лапко В. А.** Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. 2012. Т. 49. № 3.1. С. 152—156.

7. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. V. 33. № 3. P. 1065—1076.

8. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. Вып. 1. С. 156—161.

9. **Горяинов В. Б. и др.** Математическая статистика: Учеб. пособие. М.: Изд-во МГТУ им. Н. Э. Баумана, 2001.

Дата принятия 13.10.2013 г.