УДК 519.2

А. В. Лапко, В. А. Лапко

НЕПАРАМЕТРИЧЕСКАЯ ОЦЕНКА СМЕСИ ПЛОТНОСТЕЙ ВЕРОЯТНОСТИ, ОСНОВАННАЯ НА ТЕХНОЛОГИИ РАЗМНОЖЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ

Исследована непараметрическая оценка смеси плотностей вероятности, синтез которой основан на технологии размножения статистических данных. Установлены условия ее асимптотической несмещенности и состоятельности. На этой основе проведено сравнение свойств предлагаемой оценки смеси плотностей вероятности с непараметрической оценкой Розенблатта–Парзена.

Ключевые слова: плотность вероятности, непараметрическая оценка, размножение данных, асимптотические свойства.

Большинство статистических методов обработки информации ориентировано на представительные обучающие выборки. Однако при решении прикладных задач часто имеется ограниченный объем наблюдений – короткая либо малая выборка, что обусловливается нестационарностью объекта исследования, высокой стоимостью и сложностью получения дополнительной информации. Получаемые на их основе решающие правила не всегда обеспечивают приемлемые результаты, так как информации малых выборок недостаточно для оценивания вероятностных характеристик изучаемых закономерностей.

Проблемы малых выборок можно разрешить с помощью технологий обработки информации, основанных на бутстреп-методах. Ниже на основе результатов аналитических исследований будет обоснована эффективность его применения при непараметрическом оценивании плотностей вероятности.

Синтез непараметрической оценки смеси плотностей вероятности, основанной на технологии бутстреп-метода. Пусть $V = \left(x^i, i = \overline{1,n}\right)$ – выборка из n независимых наблюдений случайной величины $x = \left(x_\nu, \nu = \overline{1,k}\right)$ с плотностью вероятности $p(x)$, вид которой априори неизвестен.

Сформируем на основе исходной выборки N групп наблюдений выборку $V_j = \left(x^i, i \in I_j\right)$, где I_j – множество номеров элементов из V, составляющих j-ю группу. Количество элементов в группах одинаково и равно $\overline{n} = n - n'$. Каждая пара групп V_j, V_t, j, $t = \overline{1,N}$, $j \neq t$ отличается n' элементами. Количество групп элементов $N = n/n'$.

По каждой выборке V_j построим непараметрические оценки плотностей вероятности [1; 2]:

$$\overline{p}_j(x) = \frac{1}{\overline{n}\prod\limits_{\nu=1}^{k} c_\nu} \sum_{i \in I_j} \prod_{\nu=1}^{k} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right), \quad j = \overline{1,N}, \tag{1}$$

где $\Phi(u)$ – ядерные функции, удовлетворяющие условиям H:

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

$c_\nu = c_\nu(\overline{n})$ – коэффициенты размытости ядерных функций, значения которых убывают с ростом $\overline{n}$. Здесь и далее бесконечные пределы интегрирования опускаются.

В качестве приближения $p(x)$ по статистической выборке V примем смесь непараметрических оценок $\overline{p}_j(x)$ плотности вероятности

$$\overline{\overline{p}}(x) = \frac{1}{N}\sum_{j=1}^{N} \overline{p}_j(x). \tag{2}$$

Статистика (2) построена в соответствии с бутстреп-методом и допускает использование технологии параллельных вычислений.

Исследуем асимптотические свойства оценки плотности вероятности (2) в условиях, когда $k = 1$.

Теорема. Пусть $p(x)$ и первые две ее производные ограничены и непрерывны; ядерные функции $\Phi(u)$ удовлетворяют условиям нормированности, положительности и симметричности H; последовательность $c(\overline{n}) = c$ коэффициентов размытости ядерных функций такова, что при $\overline{n} \to \infty$ значения $c \to 0$, а при $\overline{n}c \to \infty$ и $\frac{1}{\overline{n}} \to 0$, $\frac{n'}{\overline{n}^2} \to 0$. Тогда при конечных значениях N непараметрическая оценка $\overline{\overline{p}}(x)$ смеси плотности вероятности $p(x)$ обладает свойством асимптотической несмещенности и состоятельности.

Доказательство. По определению

$$M\left(\overline{\overline{p}}(x)\right) = \frac{1}{N} \times$$
$$\times \sum_{j=1}^{N} M\left(\overline{p}_j(x)\right) = \frac{1}{c}\int \Phi\left(\frac{x-t}{c}\right) p(t)dt =$$
$$= \int \Phi(u)p(x - cu)du,$$

где M – знак математического ожидания.

Разлагая $p(x - cu)$ в ряд Тейлора и ограничиваясь первыми двумя членами ряда при $\overline{n} \to \infty$, имеем

$$W_1 = M\left(\overline{\overline{p}}(x) - p(x)\right) \sim \frac{p^{(2)}(x)}{2}c^2, \tag{3}$$

где $p^{(2)}(x)$ – вторая производная плотности вероятности $p(x)$ по x. Отсюда из условия $c = c(\overline{n})$ и $c \to 0$ при $\overline{n} \to \infty$ следует свойство асимптотической несмещенности непараметрической оценки смеси плотностей вероятности (2).

Для доказательства сходимости $\overline{\overline{p}}(x)$ в среднеквадратическом отклонении рассмотрим выражение

$$M\int\left(p(x)-\bar{\bar{p}}(x)\right)^2dx=$$
$$=M\int\left[\frac{1}{N}\sum_{j=1}^{N}\left(p(x)-\bar{p}_j(x)\right)\right]^2dx=$$
$$=\frac{1}{N^2}M\left[\begin{array}{l}\sum\limits_{j=1}^{N}\int\left(p(x)-\bar{p}_j(x)\right)^2dx+\\+\sum\limits_{j=1}^{N}\sum\limits_{\substack{t=1\\t\neq j}}^{N}\int\left(p(x)-\bar{p}_j(x)\right)\times\\\times\left(p(x)-\bar{p}_t(x)\right)dx\end{array}\right]. \quad (4)$$

Найдем асимптотическое выражение функционала

$$M\int\left(p(x)-\bar{p}_j(x)\right)\left(p(x)-\bar{p}_t(x)\right)dx=$$
$$=\int p^2(x)dx-M\int\bar{p}_t(x)p(x)dx-$$
$$-M\int\bar{p}_j(x)p(x)dx+M\int\bar{p}_j(x)\bar{p}_t(x)dx. \quad (5)$$

Преобразуем его последнее слагаемое:

$$M\int\bar{p}_j(x)\bar{p}_t(x)dx=$$
$$=\frac{1}{\bar{n}^2c^2}\int\left[\sum_{i\in I_{jt}}M\Phi^2\left(\frac{x-x^i}{c}\right)+\right.$$
$$\left.+\sum_{i\in I_j}\sum_{\nu\in I_t\setminus I_{jt}}M\Phi\left(\frac{x-x^i}{c}\right)M\Phi\left(\frac{x-x^\nu}{c}\right)\right]dx,$$

которое при достаточно большом объеме $\bar{n}$ элементов в группах V_j, $j=\overline{1,N}$, может быть представлено в виде

$$\frac{\bar{n}-n'}{\bar{n}^2c}\left\|\Phi(u)\right\|^2+\frac{\bar{n}^2-(\bar{n}-n')}{\bar{n}^2}\times$$
$$\times\int\left(p(x)+c^2p^{(2)}(x)/2\right)^2dx \quad , \quad (6)$$

где $\left\|\Phi(u)\right\|^2=\int\Phi^2(u)du$.

Заметим, что при $\bar{n}\to\infty$

$$M\int\bar{p}_t(x)p(x)dx\sim\left\|p(x)\right\|^2+\frac{c^2}{2}\int p^{(2)}(x)p(x)dx,$$

где $\left\|p(x)\right\|^2=\int p^2(x)dx$. Тогда асимптотическое выражение для функционала (5) соответствует выражению

$$\frac{\bar{n}-n'}{\bar{n}^2c}\left\|\Phi(u)\right\|^2+\frac{c^4}{4}\left\|p^{(2)}(x)\right\|^2+$$
$$+\frac{\bar{n}-n'}{\bar{n}^2}\int\left(p(x)+c^2p^{(2)}(x)/2\right)^2dx. \quad (7)$$

С учетом (7) и справедливости при $\bar{n}\to\infty$ утверждения [2]

$$M\int\left(p(x)-\bar{p}_j(x)\right)^2dx\sim\frac{\left\|\Phi(u)\right\|^2}{\bar{n}c}+\frac{c^4\left\|p^{(2)}(x)\right\|^2}{4},$$

запишем асимптотическое выражение для (4):

$$M\int\left(p(x)-\bar{\bar{p}}(x)\right)^2dx\sim\frac{1}{N}\times$$
$$\times\left(\frac{\left\|\Phi(u)\right\|^2}{\bar{n}c}+\frac{c^4\left\|p^{(2)}(x)\right\|^2}{4}\right)+$$
$$+\frac{N-1}{N}\left[\begin{array}{l}\frac{\bar{n}-n'}{\bar{n}^2c}\left\|\Phi(u)\right\|^2+\frac{c^4\left\|p^{(2)}(x)\right\|^2}{4}+\\+\frac{\bar{n}-n'}{\bar{n}^2}\int\left(p(x)+c^2p^{(2)}(x)/2\right)^2dx\end{array}\right].$$

Отсюда, пренебрегая величинами малости $0\left(\frac{1}{\bar{n}}\right)$, $0\left(\frac{n'}{\bar{n}^2}\right)$, получим

$$M\int\left(p(x)-\bar{\bar{p}}(x)\right)^2dx\sim\frac{\left\|\Phi(u)\right\|^2}{N\bar{n}c}\times$$
$$\times\left(1+\frac{(\bar{n}-n')(N-1)}{\bar{n}}\right)+\frac{c^4\left\|p^{(2)}(x)\right\|^2}{4}. \quad (8)$$

Нетрудно заметить, что при выполнении условий $c\to0$, $\bar{n}c\to\infty$ при $\bar{n}\to\infty$ оценка плотности вероятности (2) сходится в среднеквадратическом отклонении к $p(x)$, а с учетом свойства ее асимптотической несмещенности является состоятельной.

Сравнение асимптотических свойств статистики (2) и непараметрической оценки Розенблатта–Парзена. Определим минимальное значение W_2 выражения (8) при оптимальных значениях $\bar{c}$ коэффициентов размытости ядерных функций непараметрических оценок $\bar{p}_j(x)$, $j=\overline{1,N}$, составляющих их смесь $\bar{\bar{p}}(x)$ (2).

В принятых допущениях значение

$$\bar{c}=\left(\frac{\left\|\Phi(u)\right\|^2}{\bar{n}\left\|p^{(2)}(x)\right\|^2}\right)^{\frac{1}{5}}.$$

Тогда

$$W_2=\left[\left(\frac{\left\|\Phi(u)\right\|^2}{n-n'}\right)^4\left\|p^{(2)}(x)\right\|^2\right]^{\frac{1}{5}}\times$$
$$\times\left(\frac{1}{N}\left(1+\frac{n-n'}{n}(N-1)\right)+\frac{1}{4}\right)=$$
$$=\left[\left(\frac{\left\|\Phi(u)\right\|^2}{n-n'}\right)^4\left\|p^{(2)}(x)\right\|^2\right]^{\frac{1}{5}}\left(\frac{n-n'}{n}+\frac{1}{4}\right). \quad (9)$$

Если $n'=0$, то W_2 совпадает с минимальным значением асимптотического выражения среднеквадратического отклонения

$$W_2'=\frac{5}{4}\left[\left(\frac{\left\|\Phi(u)\right\|^2}{n}\right)^4\left\|p^{(2)}(x)\right\|^2\right]^{\frac{1}{5}}$$

для оценки плотности вероятности типа Розенблатта–Парзена

$$\tilde{p}(x)=\frac{1}{nc}\sum_{i=1}^{n}\Phi\left(\frac{x-x^i}{c}\right) \quad (10)$$

при оптимальных значениях $c=c^*$. При этом

$$\frac{W_2}{W_2'}=\frac{4}{5}\left(\frac{n}{n-n'}\right)^{\frac{4}{5}}\left(\frac{n-n'}{n}+\frac{1}{4}\right). \quad (11)$$

По аналогии сравним главные дисперсионные составляющие

$$W_3=\frac{n-n'}{n}\left[\left(\frac{\left\|\Phi(u)\right\|^2}{n-n'}\right)^4\left\|p^{(2)}(x)\right\|^2\right]^{\frac{1}{5}},$$

$$W_3'=\left[\left(\frac{\left\|\Phi(u)\right\|^2}{n}\right)^4\left\|p^{(2)}(x)\right\|^2\right]^{\frac{1}{5}}$$

непараметрических оценок плотностей вероятностей (2), (10). Их отношение будет следующим

$$\frac{W_3}{W_3'} = \left(\frac{n - n'}{n} \right)^{\frac{1}{5}}.$$

Так как объем статистических данных при синтезе составляющих (1) статистики (2) меньше, чем при формировании непараметрической оценки плотности вероятности (10), то очевидно, что ее смещение меньше по сравнению с оценкой плотности вероятности (2). Действительно, отношение их минимальных асимптотических выражений смещений при оптимальных значениях коэффициентов размытости имеет вид

$$\frac{W_1}{W_1'} = \left(\frac{n}{n - n'} \right)^{\frac{2}{5}} > 1. \qquad (13)$$

Значения отношений (11), (12), (13) при $n' = \alpha n$ приведены в таблице.

Использование статистики (2) позволяет несколько улучшить эффективность оценивания плотности вероятности по сравнению с оценкой Розенблатта–Парзена (10).

С уменьшением $n' = \alpha n$ растет количество групп наблюдений V_j, $j = \overline{1, N}$, но снижается уровень их разнообразия. При этом состав групп наблюдений незначительно отличается. Поэтому в данных условиях аппроксимационные свойства статистик (2) и (10) практически одинаковы.

При увеличении n' уменьшается количество составляющих смеси непараметрических оценок плотностей вероятности (2), что приводит к снижению ее аппроксимационных свойств, несмотря на рост разнообразия групп наблюдений V_j, $j = \overline{1, N}$. При этом уменьшение дисперсии смеси (2) объясняется различными темпами изменения значений ее смещения и среднеквадратического отклонения от искомой плотности вероятности.

Таким образом, использование технологии бутстреп-метода позволяет повысить эффективность оценивания плотностей вероятности. Получаемые при этом непараметрические оценки смеси плотностей вероятности обладают повышенными аппроксимационными свойствами, что особенно наблюдается в снижении их дисперсии и среднеквадратического отклонения. Определены условия их асимптотической сходимости и преимущества перед традиционной непараметрической оценкой Розенблатта–Парзена. Следует ожидать более значительного преимущества предлагаемой методики оценивания плотностей вероятности в условиях малых выборок.

Библиографический список

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

2. Епанечников, В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. Вып. 1. С. 156–161.

Зависимость отношений $W_i / W_i', i = \overline{1, 3}$ от значений n'

$\frac{W_i}{W_i'}$	α						
	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{20}$	$\frac{1}{40}$
(11)	1	1,044	1,002	0,998	0,999	0,999	0,999
(12)	1	0,87	0,94	0,97	0,987	0,989	0,995
(13)	1	1,32	1,12	1,054	1,024	1,020	1,008

A. V. Lapko, V. A. Lapko

NONPARAMETRIC ESTIMATION OF A MIX PROBABILITY DENSITY, BASED ON TECHNOLOGY DUPLICATION OF STATISTICAL DATA

The nonparametric estimation of the mix probability density which synthesis is based on technology duplication of statistical data is investigated. The asymptotic unbiasedness conditions and solvencies are determined. On this basis the properties comparison of the offered estimation of the mix probability density with the nonparametric estimation such as Rozenblatt–Parzen is carried out.

Keywords: probability density, nonparametric estimation, data duplication, asymptotic properties.