

АНАЛИЗ СВОЙСТВ СМЕСИ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК ПЛОТНОСТИ ВЕРОЯТНОСТИ МНОГОМЕРНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ*

Исследуются асимптотические свойства смеси непараметрических оценок плотности вероятности многомерной случайной величины. Проводится их сравнение со свойствами традиционной непараметрической оценки плотности вероятности типа Розенблатта–Парзена в зависимости от количества составляющих смеси и размерности случайной величины.

Ключевые слова: смесь плотностей вероятности, непараметрическая оценка, большие выборки, асимптотические свойства.

Использование методов непараметрической статистики, основанных на оценках плотности вероятности типа Розенблатта–Парзена [1; 2] является одним из активно развивающихся направлений моделирования систем при априорной неопределённости. Однако при усложнении условий исследования систем появляются методические и вычислительные трудности применения традиционных непараметрических алгоритмов и моделей, что особенно наблюдается при обработке статистических данных большого объёма.

Перспективное направление «обхода» возникающих проблем состоит в применении принципов декомпозиции обучающих выборок по их объёму и использовании технологии параллельных вычислений.

Цель данной работы: на основе анализа асимптотических свойств непараметрической оценки смеси плотностей вероятности обосновать эффективность использования принципов декомпозиции при обработке больших массивов статистических данных.

Непараметрическая оценка смеси плотностей вероятности и её свойства. Пусть выборка $V = (x^i, i = 1, n)$ из n независимых наблюдений k -мерной случайной величины $x = (x_v, v = 1, k)$ с плотностью вероятности $p(x)$. Вид $p(x)$ априори неизвестен.

Разобьём выборку V на T групп наблюдений $V_j = (x^i, i \in I_j)$, $j = 1, T$. Множество номеров наблюдений x в группе под номером j обозначим через I_j .

Причем $\bigcup_{j=1}^T I_j = I = (i = 1, n)$. Количество $n_j = |I_j|$ элементов в выборках $V_j = (x^i, i \in I_j)$ одинаково и равно $\bar{n} = \frac{n}{T}$.

По каждой выборке V_j построим непараметрическую оценку плотности вероятности многомерной случайной величины x [1]:

$$\bar{p}_j(x) = \frac{1}{\bar{n} \prod_{v=1}^k c_v} \sum_{i \in I_j} \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v}\right), \quad j = 1, T. \quad (1)$$

В статистике (1) ядерные функции $\Phi(u_v)$, удовлетворяют условиям нормированности, положительности и симметричности. Параметры ядерных функций $c_v = c_v(\bar{n})$ убывают с ростом $\bar{n}$.

Пусть интервалы изменения значений компонент x_v вектора x одинаковы. В этих условиях появляется возможность полагать, что значения коэффициентов c_v непараметрических оценок плотностей вероятности $\bar{p}_j(x)$, $j = 1, T$ одинаковы и равны c . Тогда непараметрическая оценка (1) плотности вероятности принимает вид

$$\bar{p}_j(x) = \frac{1}{\bar{n} c^k} \sum_{i \in I_j} \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c}\right), \quad j = 1, T. \quad (2)$$

В качестве приближения $p(x)$ по статистической выборке V будем использовать смесь непараметрических оценок плотности вероятности типа

$$\bar{\bar{p}}(x) = \frac{1}{T} \sum_{j=1}^T \bar{p}_j(x). \quad (3)$$

Статистика (3) допускает использование технологии параллельных вычислений при оценивании плотности вероятности в условиях больших выборок.

Асимптотические свойства $\bar{\bar{p}}(x)$ определяются следующим утверждением.

Теорема. Пусть $p(x)$ и первые две ее производные по каждой компоненте x_v , $v = 1, k$ ограничены и непрерывны; ядерные функции $\Phi(u_v)$ удовлетворяют условиям

$$\begin{aligned} \Phi(u_v) &= \Phi(-u_v), \quad 0 \leq \Phi(u_v) < \infty, \\ \int \Phi(u_v) du_v &= 1, \quad \int u_v^2 \Phi(u_v) du_v = 1, \\ \int u_v^m \Phi(u_v) du_v &< \infty, \quad 0 \leq m < \infty; \quad v = 1, k, \end{aligned}$$

последовательности $c = c(\bar{n})$ коэффициентов размытости ядерных функций непараметрических оценок плотности вероятности (1) таковы, что при $\bar{n} \rightarrow \infty$ значения $c \rightarrow 0$, а $\bar{n} c^k \rightarrow \infty$.

Тогда при конечных значениях T непараметрическая оценка (3) плотности вероятности $p(x)$ обладает свойством асимптотической несмещённости и состоятельности.

Здесь и далее бесконечные пределы интегрирования опускаются.

Доказательство.

По определению имеем

$$\begin{aligned} M(\bar{\bar{p}}(x)) &= \frac{1}{T} \sum_{j=1}^T M(\bar{p}_j(x)) = \frac{1}{T} \sum_{j=1}^T \frac{1}{\bar{n} c^k} \times \\ &\times \sum_{i \in I_j} \int \dots \int \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c}\right) p(x_1^i, \dots, x_k^i) dx_1^i \dots dx_k^i = \end{aligned}$$

* Работа выполнена при поддержке гранта ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг., ГК № 02.740.11.0621.

$$= \frac{1}{c^k} \int \dots \int \prod_{v=1}^k \Phi\left(\frac{x_v - t_v}{c}\right) p(t_1, \dots, t_k) dt_1 \dots dt_k =$$

$$= \int \dots \int \prod_{v=1}^k \Phi(u_v) p(x_1 - cu_1, \dots, x_k - cu_k) du_1 \dots du_k,$$

где M – знак математического ожидания. При выполнении преобразований учитывается, что элементы статистических выборок V_j , $j = \overline{1, T}$ являются значениями одной и той же случайной величины t с плотностью вероятности $p(t_1, \dots, t_k)$.

Разложим $p(x_1 - cu_1, \dots, x_k - cu_k)$ в ряд Тейлора в точке $x = x_1, \dots, x_k$ и ограничиваясь первыми двумя членами ряда, имеем

$$W_1 = M(\bar{p}(x) - p(x)) \sim \frac{c^2}{2} \sum_{v=1}^k p_v^{(2)}(x), \quad (4)$$

где $p_v^{(2)}(x)$ – вторая производная плотности вероятности $p(x)$ по компоненте x_v .

Отсюда, из условия $c \rightarrow 0$ при $\bar{n} \rightarrow \infty$ следует свойство асимптотической несмещенности смеси непараметрических оценок плотности вероятности (3).

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

$$M \int \dots \int (p(x) - \bar{p}(x))^2 dx_1 \dots dx_k =$$

$$= M \int \dots \int \left[\frac{1}{T} \sum_{j=1}^T (p(x) - \bar{p}_j(x)) \right]^2 dx_1 \dots dx_k =$$

$$= \frac{1}{T^2} M \left[\sum_{j=1}^T \int \dots \int (p(x) - \bar{p}_j(x))^2 dx_1 \dots dx_k + \right.$$

$$\left. + \sum_{j=1}^T \sum_{t \neq j}^T \int \dots \int (p(x) - \bar{p}_j(x))(p(x) - \bar{p}_t(x)) dx_1 \dots dx_k \right]. \quad (5)$$

Найдем асимптотическое выражение составляющих второй части выражения (5):

$$M \int \dots \int (p(x) - \bar{p}_j(x))(p(x) - \bar{p}_t(x)) dx_1 \dots dx_k =$$

$$= \int \dots \int p^2(x) dx_1 \dots dx_k -$$

$$- M \int \dots \int \bar{p}_t(x) p(x) dx_1 \dots dx_k - M \times$$

$$\times \int \dots \int \bar{p}_j(x) p(x) dx_1 \dots dx_k + M \int \dots \int \bar{p}_j(x) \bar{p}_t(x) dx_1 \dots dx_k. \quad (6)$$

Преобразуем его последнее слагаемое

$$M \int \dots \int \bar{p}_j(x) \bar{p}_t(x) dx_1 \dots dx_k =$$

$$= \int \dots \int M(\bar{p}_j(x)) M(\bar{p}_t(x)) dx_1 \dots dx_k,$$

которое при достаточно большом объеме статистических данных с учетом выражения (4) представляется в виде

$$\int \dots \int \left(p(x) + \frac{c^2}{2} \sum_{v=1}^k p_v^{(2)}(x) \right)^2 dx_1 \dots dx_k. \quad (7)$$

Заметим, что асимптотическое выражение статистики типа

$$M \int \dots \int \bar{p}_t(x) p(x) dx_1 \dots dx_k$$

соответствует

$$\int \dots \int \left(p(x) + \frac{c^2}{2} \sum_{v=1}^k p_v^{(2)}(x) \right) p(x) dx_1 \dots dx_k. \quad (8)$$

Подставляя выражения (7), (8) в (6), после несложных преобразований получим

$$M \int \dots \int (p(x) - \bar{p}_j(x))(p(x) - \bar{p}_t(x)) dx_1 \dots dx_k \sim$$

$$\sim \frac{c^4}{4} \int \dots \int \left(\sum_{v=1}^k p_v^{(2)}(x) \right)^2 dx_1 \dots dx_k = \frac{c^4}{4} B. \quad (9)$$

В работе В. А. Епанечникова [2] получено асимптотическое выражение для среднеквадратического отклонения непараметрической оценки плотности вероятности от $p(x)$, составляющих первую часть выражения (5):

$$M \int \dots \int (p(x) - \bar{p}_j(x))^2 dx_1 \dots dx_k \sim$$

$$\sim \frac{\prod_{v=1}^k \int \Phi^2(u_v) du_v}{\bar{n} c^k} + \frac{c^4}{4} B. \quad (10)$$

С учетом (9), (10) выражение (5) при достаточно больших значениях $\bar{n}$ представляется в виде

$$M \int \dots \int (p(x) - \bar{p}(x))^2 dx_1 \dots dx_k \sim \frac{\prod_{v=1}^k \int \Phi^2(u_v) du_v}{T \bar{n} c^k} + \frac{c^4}{4} B. \quad (11)$$

Нетрудно заметить, что в условиях $c \rightarrow 0$, $\bar{n} c^k \rightarrow \infty$ при $\bar{n} \rightarrow \infty$ оценка смеси плотностей вероятности (3) сходится в среднеквадратическом к $p(x)$, а с учетом свойства ее асимптотической несмещенности является состоятельной.

При $T = 1$ и $\bar{n} = n$ полученный результат (11) совпадает с утверждением теоремы В. А. Епанечникова [2], что подтверждает корректность выполненных преобразований.

Анализ аппроксимационных свойств статистики $\bar{p}(x)$. Для анализа эффективности непараметрической оценки смеси плотностей вероятности (3) и оценки плотности вероятности типа Розенблатта–Парзена

$$\tilde{p}(x) = \frac{1}{n c^k} \sum_{i=1}^n \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c}\right) \quad (12)$$

рассмотрим отношение соответствующих им асимптотических выражений среднеквадратических отклонений при оптимальных значениях коэффициентов размытости ядерных функций.

Определим минимальное значение W_2 выражения (11) при оптимальных значениях c^* коэффициентов размытости непараметрических оценок $\bar{p}_j(x)$ составляющих смеси плотностей вероятности. В принятых допущениях значение

$$c^* = \left(\frac{k \prod_{v=1}^k \int \Phi^2(u_v) du_v}{\bar{n} B} \right)^{1/(k+4)}.$$

Тогда

$$W_2 = \left[\left(\frac{\prod_{v=1}^k \int \Phi^2(u_v) du_v}{\bar{n}} \right)^4 B^k \right]^{1/(k+4)} \frac{4 + T k}{4 T k^{k/(k+4)}}. \quad (13)$$

Если $k = 1$, то W_2 совпадает с минимальным асимптотическим выражением среднеквадратического отклонения для смеси непараметрических оценок плотностей вероятности, полученного в работе [3].

При $T = 1$ и $\bar{n} = n$ выражение (13) соответствует минимальному асимптотическому выражению W'_2 среднеквадратического отклонения для непараметрической оценки плотности вероятности (12) типа Розенблатта–Парзена [2].

Отсюда, после несложных преобразований, получим

$$R_2 = \frac{W_2}{W'_2} = \frac{4 + T k}{(4 + k) T^{k/(k+4)}}.$$

По аналогии вычислим отношение минимальных значений главных дисперсионных составляющих статистик $\bar{\bar{p}}(x)$ и $\tilde{p}(x)$:

$$W_3 = \frac{1}{T k^{k/(k+4)}} \left[\left(\frac{\prod_{v=1}^k \int \Phi^2(u_v) du_v}{\bar{n}} \right)^4 B^k \right]^{1/(k+4)},$$

$$W'_3 = \frac{1}{k^{k/(k+4)}} \left[\left(\frac{\prod_{v=1}^k \int \Phi^2(u_v) du_v}{n} \right)^4 B^k \right]^{1/(k+4)}.$$

Их отношение имеет вид

$$R_3 = \frac{W_3}{W'_3} = \frac{1}{T^{k/(k+4)}}.$$

Нетрудно убедиться, что отношение асимптотических выражений смещений W_1, W'_1 анализируемых оценок плотности вероятности $\bar{\bar{p}}(x)$ и $\tilde{p}(x)$ при оптимальных коэффициентах размытости ядерных функций соответствует значению

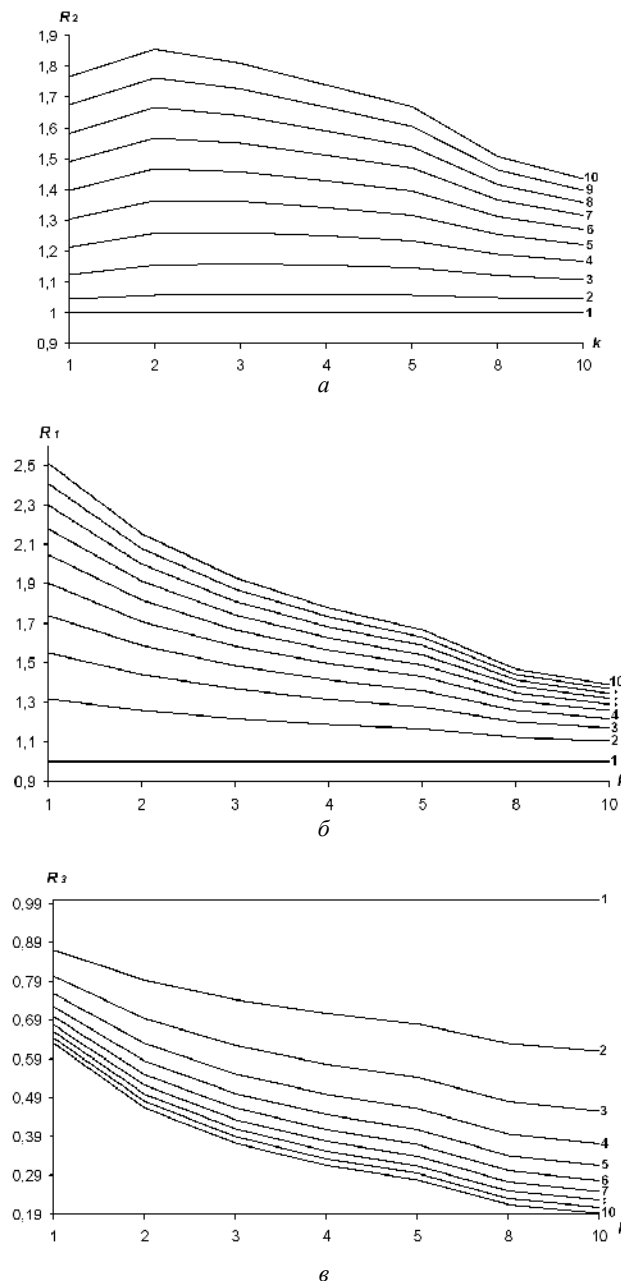
$$R_1 = \frac{W_1}{W'_1} = T^{2/(k+4)}.$$

С ростом количества T составляющих смеси непараметрических оценок плотности вероятности наблюдается увеличение значений отношений $R_2 > 1$ (см. рисунок, часть а), $R_1 > 1$ (см. рисунок, часть б). Отмеченное ухудшение аппроксимационных свойств смеси $\bar{\bar{p}}(x)$ по сравнению с традиционной непараметрической оценкой плотности вероятности $\tilde{p}(x)$ (12) объясняется снижением объёмов выборок, используемых при оценивании составляющих $\bar{\bar{p}}(x)$. Данная тенденция особенно характерна для малых размерностей k случайной величины x . При усложнении условий оценивания плотности вероятности с ростом k эффективность непараметрических оценок $\bar{\bar{p}}(x)$ и $\tilde{p}(x)$ снижается. Соответствующие им критерии W_2, W'_2 и W_1, W'_1 становятся соизмеримыми, что проявляется в снижении значений их отношений R_2 и R_1 .

Предлагаемая смесь $\bar{\bar{p}}(x)$ оценок плотности вероятности обладает меньшей дисперсией по сравнению с непараметрической оценкой $\tilde{p}(x)$, что обусловлено её структурой, так как синтез статистики $\bar{\bar{p}}(x)$ осуществляется на основе усредняющего оператора (см. рисунок, часть в). Причём с увеличением количества T составляющих смеси непараметрических оценок $\bar{\bar{p}}(x)$ плотности вероятности и размерности k случайной величины её преимущество возрастает.

На основе анализа асимптотических свойств смеси непараметрических оценок плотности вероятности мно-

гомерной случайной величины обоснована возможность декомпозиции исходных статистических данных при синтезе непараметрических статистик в условиях больших выборок. Исследуемая статистика по сравнению с традиционной непараметрической оценкой плотности вероятности типа Розенблатта–Парзена имеет значительно меньшую дисперсию и позволяет использовать технологию параллельных вычислений.



Зависимости отношений: R_2 (а), R_1 (б), R_3 (в) от размерности k случайной величины $x = (x_v, v = \overline{1, k})$ и количества $T = 1-10$ (кривые 1, ..., 10) составляющих смеси непараметрических оценок плотности вероятности $\bar{\bar{p}}(x)$ (3)

Библиографические ссылки

1. Parzen E. On estimation of a probability density function and mode // 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.

3. Лапко В. А., Варочкин С. С., Егорочкин И. А. Разработка и исследование непараметрической оценки плот-

ности вероятности, основанной на принципе декомпозиции обучающей выборки по её объёму // Вестник СибГАУ. 2009. № 1(22). Ч. 1. С. 45–49.

A. V. Lapko, V. A. Lapko

ANALYSIS OF PROPERTIES OF MIXTURE OF NONPARAMETRIC ESTIMATIONS OF A PROBABILITY DENSITY OF A MULTIDIMENSIONAL RANDOM VARIABLE

Asymptotic properties of mixture of nonparametric estimations of a probability density of a multidimensional random variable are researched. Their correlation with properties of a traditional nonparametric estimation of a probability density of Rosenblatt–Parzen type, in accordance with quantity of components of mixture and dimension of a random variable is arranged.

Keywords: mixture of probability densities, nonparametric estimation, the big samples, asymptotic properties.

© Лапко А. В., Лапко В. А., 2010

УДК 681.513

А. В. Лапко, В. А. Лапко

СВОЙСТВА НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ УРАВНЕНИЯ РАЗДЕЛЯЮЩЕЙ ПОВЕРХНОСТИ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ ПРИ СЛУЧАЙНЫХ ЗНАЧЕНИЯХ КОЭФФИЦИЕНТОВ РАЗМЫТОСТИ ЯДЕРНЫХ ФУНКЦИЙ*

Исследуются асимптотические свойства непараметрической оценки уравнения разделяющей поверхности, основанной на рандомизированном методе её оптимизации. Проводится их сравнение со свойствами традиционной непараметрической решающей функции парзеновского типа.

Ключевые слова: непараметрическая статистика, распознавание образов, случайные коэффициенты размытости, асимптотические свойства.

Существующий парадокс традиционных методов идентификации стохастических моделей состоит в сопоставлении конечной случайной выборке наблюдений переменных изучаемых объектов конкретного набора параметров модели, оптимальных в некотором смысле.

Впервые возможность случайного выбора коэффициентов размытости ядерных функций при синтезе непараметрической оценки плотности вероятности типа Розенблатта–Парзена была реализована в 1975 г. Т. Вагнером [1]. В работе [2] была предложена методика синтеза непараметрических алгоритмов распознавания образов, основанная на рандомизированном методе её оптимизации. Её идея состоит в признании случайного характера коэффициентов размытости ядерных функций в условиях обучающей выборки конечного объёма и выборе параметров закона их распределения при оптимизации непараметрических решающих правил. На основе анализа асимптотических свойств непараметрической оценки

плотности вероятности типа Розенблатта–Парзена со случайными коэффициентами размытости ядерных функций показана возможность нахождения рационального закона распределения в классе степенных функций. Однако исследование соответствующих непараметрических алгоритмов распознавания образов осуществлялось по данным вычислительных экспериментов.

Цель данной работы состоит в установлении асимптотических свойств непараметрической оценки уравнения разделяющей поверхности, основанной на рандомизированном методе её оптимизации, и их количественной зависимости от параметров закона распределения коэффициентов размытости ядерных функций.

Непараметрический алгоритм распознавания образов со случайными коэффициентами размытости ядерных функций. Рассмотрим методику построения непараметрического классификатора на примере двальтернативной задачи распознавания образов в пространстве непрерывного признака x .

* Работа выполнена при поддержке гранта ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг., ГК № 02.740.11.0621.