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 плотности вероятности, основанной на принципе декомпозиции обучающей выборки по её объему // 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.

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 двуальтернативной 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.

Известно, что байесовское решающее правило распознавания образов, соответствующее критерию максимального правдоподобия, имеет вид [3]

$$m(x):\begin{cases} x\in\Omega_1, \text{ если } f_{12}(x)\le 0 \\ x\in\Omega_2, \text{ если } f_{12}(x)>0. \end{cases} \quad (1)$$

Здесь

$$f_{12}(x)=p_2(x)-p_1(x) \quad (2)$$

является уравнением разделяющей поверхности между классами Ω_1, Ω_2; $p_j(x)$ – условная плотность вероятности распределения признака x анализируемых объектов в классе Ω_j, $j=1,2$.

В условиях априорной неопределённости о виде законов распределения $p_j(x), j=1,2$ при статистическом оценивании уравнения разделяющей поверхности используются непараметрические методы статистики.

Пусть $V=\left(x^i,\sigma(i), i=\overline{1,n}\right)$ – обучающая выборка объёма n, составленная из значений признака x^i классифицируемого объекта и соответствующего ему «указания учителя» $\sigma(i)$ о его принадлежности к одному из двух классов Ω_1, Ω_2.

Для оценивания плотности вероятности в уравнении разделяющей поверхности (2) будем использовать статистику типа Розенблатта–Парзена [4].

Тогда традиционная непараметрическая оценка уравнения разделяющей поверхности $f_{12}(x)$ (2) представляется в виде

$$\tilde{f}_{12}(x)=(n\,c)^{-1}\sum_{i=1}^{n}\sigma_1(i)\Phi\left(\frac{x-x^i}{c}\right). \quad (3)$$

Здесь

$$\sigma_1(i)=\begin{cases} -\overline{P}_1^{-1}, \text{ если } x^i\in\Omega_1, \\ \overline{P}_2^{-1}, \text{ если } x^i\in\Omega_2; \end{cases}$$

$\overline{P}_j=\dfrac{n_j}{n}$ – оценка априорной вероятности принадлежности ситуаций обучающей выборки к классу Ω_j, $j=1,2$. В статистике (3) ядерные функции $\Phi(u)$ удовлетворяют условиям

$$\Phi(u)=\Phi(-u),\ 0\le\Phi(u)<\infty,$$
$$\int\Phi(u)du=1,\ \int u^2\Phi(u)du=1, \quad (4)$$
$$\int u^m\Phi(u)du<\infty,\ 0\le m<\infty;$$

c – коэффициент размытости ядерных функций, значения которого убывают с ростом объёма n обучающей выборки.

Здесь и далее бесконечные пределы интегрирования опускаются.

Основываясь на результатах работы [2], будем искать рациональный закон распределения коэффициентов размытости c в непараметрической оценке уравнения разделяющей поверхности среди функций вида

$$p_h(c)=\alpha c^t,\ \alpha=\frac{t+1}{h^{t+1}}\ \ \forall\ c\in[0,h],$$

где h – правая граница интервала изменения c.

Параметр t плотности вероятности $p_h(c)$ априори не определён.

В соответствии с методикой построения датчиков случайных величин [5] определим процедуру формирования последовательности коэффициентов размытости ядерных функций в виде

$$c=h\,\varepsilon^{1/(t+1)}, \quad (5)$$

которая следует из решения уравнения

$$\varepsilon=\int_0^c p_h(u)\,du,$$

где $\varepsilon\in[0;1]$ – случайная величина с равномерным законом распределения.

На основании процедуры (5) сформируем последовательность $c^i, i=\overline{1,n}$ коэффициентов размытости и сопоставим случайным образом её элементам ядерные функции в статистике (3). Тогда непараметрическая оценка уравнения разделяющей поверхности со случайными коэффициентами размытости ядерных функций для двуальтернативной задачи распознавания образов запишется в виде

$$\overline{f}_{12}(x)=\frac{1}{n}\sum_{i=1}^{n}\sigma_1(i)\frac{1}{c^i}\Phi\left(\frac{x-x^i}{c^i}\right). \quad (6)$$

Оптимизация непараметрических решающих функций $\tilde{f}_{12}(x)$, $\overline{f}_{12}(x)$ осуществляется соответственно по коэффициенту размытости c и параметру h в режиме «скользящего экзамена» из условия минимума оценки вероятности ошибки распознавания образов.

Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности. Асимптотические свойства статистики (6) определяются следующим утверждением.

Теорема. Пусть плотности вероятности $p_j(x), j=1,2$ распределения x в классах и первые две их производные ограничены и непрерывны; ядерные функции $\Phi(u)$ удовлетворяют условиям нормированности, положительности и симметричности (4); последовательность $h(n)=h$ правой границы области определения плотности вероятности $p_h(c)$ коэффициентов размытости c ядерных функций таковы, что при $n_1\to\infty$, $n_2\to\infty$, значения $h\to 0$, а $\dfrac{n_1+n_2}{n_1 n_2 h}\to 0$. Тогда непараметрическая оценка $\overline{f}_{12}(x)$ уравнения разделяющей поверхности $f_{12}(x)$ обладает свойствами асимптотической несмещенности и состоятельности.

Доказательство.

1. По определению имеем

$$M\left(\overline{f}_{12}(x)\right)=M\left(\overline{p}_2(x)-\overline{p}_1(x)\right)=$$
$$=\int_0^h\left[\frac{1}{c}\int\Phi\left(\frac{x-t}{c}\right)p_2(t)dt-\frac{1}{c}\int\Phi\left(\frac{x-t}{c}\right)p_1(t)dt\right]p_h(c)\,dc,$$

где M – знак математического ожидания.

При выполнении данных преобразований учитывается, что элементы статистических выборок, определяющих каждый класс, являются значениями одной и той же случайной величины t с плотностью вероятности $p_j(t)$, $j=\overline{1,2}$. Причем элементы последовательности $c^i, i=\overline{1,n}$ формируются в соответствии с плотностью вероятности $p_h(c)$.

Проведём в интегралах последнего выражения замену переменных $(x-t)c^{-1}=u$. Разложим функции $p_j(x-cu), j=1,2$ в ряд Тейлора в точке x с учётом свойств ядерной функции (4). Тогда при достаточно больших значениях n_1, n_2 получим

$$W_1(h) = M\left(\overline{f}_{12}(x) - f_{12}(x)\right) \sim$$
$$\sim \frac{1}{2}\left(p_2^{(2)}(x) - p_1^{(2)}(x)\right)\int_0^h c^2 p_h(c)\,dc =$$
$$= \frac{(t+1)\,h^2}{t+3}\left(p_2^{(2)}(x) - p_1^{(2)}(x)\right), \qquad (7)$$

где $p_j^{(2)}(x)$ – вторая производная плотности вероятности $p_j(x)$ по x, $j = 1, 2$.

Отсюда, из условия $h \to 0$ при $n_1 \to \infty, n_2 \to \infty$ следует свойство асимптотической несмещенности непараметрической статистики $\overline{f}_{12}(x)$ (6).

2. Для доказательства состоятельности непараметрической оценки $\overline{f}_{12}(x)$ исследуем асимптотические свойства среднеквадратического отклонения

$$W_2(h) = \int_0^h \left(M\int\left(\overline{f}_{12}(x) - f_{12}(x)\right)^2 dx\right) p_h(c)\,dc. \quad (8)$$

Преобразуем выражение

$$M\int\left(\overline{f}_{12}(x) - f_{12}(x)\right)^2 dx = M\int\left(p_1(x) - \overline{p}_1(x)\right)^2 dx -$$
$$-2M\int\left(p_1(x) - \overline{p}_1(x)\right)\left(p_2(x) - \overline{p}_2(x)\right)dx +$$
$$+M\int\left(p_2(x) - \overline{p}_2(x)\right)^2 dx. \qquad (9)$$

Известно [6], что в асимптотике среднеквадратическое отклонение $\overline{p}_j(x)$ от $p_j(x)$ имеет вид

$$M\int\left(p_j(x) - \overline{p}_j(x)\right)^2 dx \sim \frac{\left\|\Phi(u)\right\|^2}{n_j c} + \frac{c^4\left\|p_j^{(2)}(x)\right\|^2}{4},$$
$$j = 1, 2,$$

а асимптотическое выражение смещения

$$M\left(p_j(x) - \overline{p}_j(x)\right) \sim \frac{c^2}{2}p_j^{(2)}(x),$$

где $\left\|\Phi(u)\right\|^2 = \int \Phi^2(u)\,du$, $\left\|p_j^{(2)}(x)\right\|^2 = \int\left(p_j^{(2)}(x)\right)^2 dx$.

Тогда при достаточно больших значения n_1, n_2 асимптотическое выражение для среднеквадратического отклонения (9) представляется в виде

$$M\int\left(\overline{f}_{12}(x) - f_{12}(x)\right)^2 dx \sim \frac{\left\|\Phi(u)\right\|^2(n_1+n_2)}{n_1 n_2 c} + \frac{c^4}{4}B, \quad (10)$$

где $B = \left\|p_1^{(2)}(x) - p_2^{(2)}(x)\right\|^2$.

С учётом (10) вычислим выражение (8):

$$W_2(h) \sim \frac{\left\|\Phi(u)\right\|^2(n_1+n_2)(t+1)}{n_1 n_2 h t} + \frac{h^4(t+1)}{4(t+5)}B. \quad (11)$$

Нетрудно заметить, что при выполнении условий $h \to 0, \frac{n_1+n_2}{n_1 n_2 h} \to 0$ при $n_j \to \infty$, $j = 1, 2$ непараметрическая оценка $\overline{f}_{12}(x)$ сходится в среднеквадратическом к байесовскому уравнению разделяющей поверхности (2), а с учётом свойства её асимптотической несмещённости является состоятельной оценкой.

Сравнение свойств непараметрических оценок уравнения разделяющей поверхности. Для анализа эффективности непараметрических оценок $\tilde{f}_{12}(x)$ и $\overline{f}_{12}(x)$ рассмотрим отношения соответствующих им асимптотических выражений среднеквадратических отклонений при оптимальных значениях параметров h и c. Определим минимальное значение $W_2(h^*)$ при оптимальном значении h^* правой границы h плотности вероятности $p_h(c)$.

Из условия минимума $W_2(h)$ получим

$$h^* = \left[\frac{\left\|\Phi(u)\right\|^2(n_1+n_2)(t+5)}{n_1 n_2 t\,B}\right]^{\frac{1}{5}} = c^*\left(\frac{t+5}{t}\right)^{\frac{1}{5}},$$

где c^* – оптимальный коэффициент размытости ядерных функций непараметрической оценки уравнения разделяющей поверхности (3) в смысле минимума среднеквадратического отклонения $\tilde{f}_{12}(x)$ от $f_{12}(x)$ при $n \to \infty$.

Тогда минимальное значение асимптотического выражения среднеквадратического отклонения $\overline{f}_{12}(x)$ от $f_{12}(x)$ запишется в виде

$$W_2(h^*) = \frac{t+1}{\left(t^4(t+5)\right)^{1/5}}\left[\frac{\left\|\Phi(u)\right\|^2(n_1+n_2)}{n_1 n_2 c^*} + \frac{(c^*)^4}{4}B\right] =$$
$$= \frac{t+1}{\left(t^4(t+5)\right)^{1/5}}W_2(c^*),$$

где $W_2(c^*)$ – минимальное значение асимптотического выражения среднеквадратического отклонения $\tilde{f}_{12}(x)$ от $f_{12}(x)$.

Тогда отношение

$$R_2 = \frac{W_2(c^*)}{W_2(h^*)} = \frac{\left(t^4(t+5)\right)^{1/5}}{t+1}. \qquad (12)$$

Сравним главные дисперсионные составляющие $W_3(h), W_3(c)$ статистик $\overline{f}_{12}(x)$, $\tilde{f}_{12}(x)$, которые соответствуют первым слагаемым выражений (11), (10) при оптимальных значениях h и c. Можно показать, что их отношение совпадает с выражением (12).

По аналогии определим отношение минимальных значений смещений для $\overline{f}_{12}(x)$, $\tilde{f}_{12}(x)$, которые определяются выражениями (7) и

$$W_1(c) = \frac{c^2}{2}\left(p_2^{(2)}(x) - p_1^{(2)}(x)\right).$$

После несложных преобразований получим

$$R_1 = \frac{W_1(c^*)}{W_1(h^*)} = \frac{t+3}{t+1}\left(\frac{t}{t+5}\right)^{\frac{2}{5}}.$$

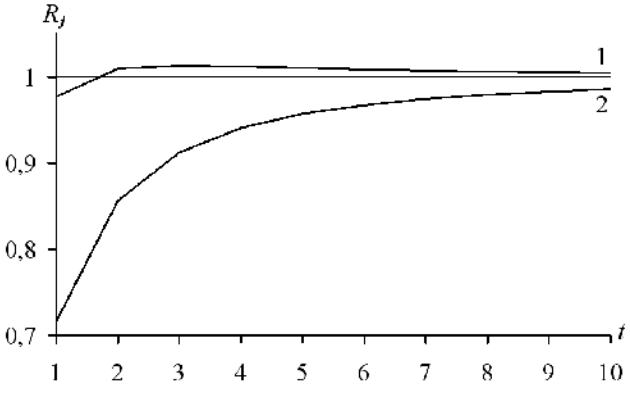
Полученные отношения R_1 и R_2 совпадают с результатами исследований непараметрической оценки плотности вероятности со случайными коэффициентами размытости ядерных функций [2]. Это объясняется тем, что статистика $\overline{f}_{12}(x)$ является линейным функционалом непараметрических оценок плотности вероятности $\overline{p}_j(x)$, $j = 1, 2$.

Предлагаемая непараметрическая оценка уравнения разделяющей поверхности $\overline{f}_{12}(x)$ имеет меньшее смещение по сравнению с традиционной непараметрической решающей функцией $\tilde{f}_{12}(x)$ (см. рисунок). С увеличением параметра t плотности вероятности $p_h(c)$ значения отношений R_1 и R_2 стремятся к 1.

В рамках предложенного подхода появляется возможность снижения значений дисперсии статистики $\overline{f}_{12}(x)$ и её среднеквадратического отклонения путём построения коллектива непараметрических оценок уравнения разделяющей поверхности между классами:

$$\overline{\overline{f}}_{12}(x) = \frac{1}{N}\sum_{j=1}^{N}\overline{f}_{12}^{\,j}(x).$$

Его составляющие характеризуются одним и тем же оптимальным параметром h правой границы области определения $p_h(c)$, но разными случайными последовательностями коэффициентов размытости ядерных функций $\left(c_j^i,\ i=\overline{1,n}\right),\ j=\overline{1,N}$.



Зависимости отношений R_1 (кривая 1), R_2, R_3 (кривая 2) от параметра t закона распределения $p_h(c)$ коэффициента размытости ядерных функций непараметрической оценки уравнения разделяющей поверхности $\bar{f}_{12}(x)$

Таким образом, непараметрическая оценка уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов, основанная на рандомизированном методе её оптимизации, обладает свойствами асимптотической несмещённости и состоятельности. По сравнению с традиционной непараметрической решающей функцией парзеновского типа предлагаемая статистика имеет меньшее смещение, но большее значение среднеквадратического отклонения.

Перспективность данного направления исследований состоит в возможности использования принципов коллективного оценивания для повышения аппроксимационных свойств непараметрических оценок уравнения разделяющей поверхности и создания алгоритмических средств их доверительного оценивания.

Библиографические ссылки

1. Деврой Л., Дьерди Л. Непараметрическое оценивание плотности (L_1 -подход). М. : Мир, 1988.

2. Лапко А. В., Лапко В. А. Непараметрические алгоритмы распознавания образов при случайных значениях коэффициентов размытости ядерных функций // Автометрия. 2007. № 5. С. 47–55.

3. Цыпкин Я. З. Основы теории обучающихся систем. М. : Наука, 1970.

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

5. Бусленко Н. П., Шрейдер Ю. А. Метод статистических испытаний. М. : Гос. изд-во физ.-мат. лит., 1961.

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

A. V. Lapko, V. A. Lapko

PROPERTIES OF A NONPARAMETRIC ESTIMATION OF THE EQUATION OF A SEPARATING SURFACE IN THE PATTERN RECOGNITION TASK AT CASUAL VALUES OF FUZZINESS COEFFICIENTS OF KERNEL FUNCTIONS

Asymptotic properties of nonparametric estimation of the equation of the separating surface grounded on the randomized method of the estimation optimization are researched. Their correlation with properties of traditional nonparametric decision function of type a Rosenblatt–Parzen is made.

Keywords: nonparametric statistics, pattern recognition, random fuzziness parameters, asymptotic properties.