УДК 519.7

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ГРАНИЦ ДЛЯ РЕШАЮЩЕЙ ФУНКЦИИ В ДВУАЛЬТЕРНАТИВНОЙ ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ*

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

Рассматривается непараметрическая оценка решающей функции в двуальтернативной задаче распознавания образов. При её синтезе используются принцип декомпозиции обучающей выборки и анализ вероятностных характеристик получаемых множеств случайных величин. На этой основе разработана методика построения доверительных границ для байесовского уравнения разделяющей поверхности. Эффективность методики подтверждается результатами вычислительных экспериментов.

Ключевые слова: распознавание образов, решающая функция, непараметрическая оценка, доверительное оценивание, правило Хайнкольда — Гаеде.

Введение. Доверительное оценивание плотности вероятности решающих функций в задачах распознавания образов и восстановления стохастических зависимостей имеет важное значение при оценивании эффективности процедур обработки информации в условиях априорной неопределённости. Вместе с тем решение данных проблем математической статистики находится на стадии становления. Наиболее широко распространён метод построения доверительных границ плотности вероятности на основе гистограммы и критерия Пирсона [1, 2]. В работе [3] впервые показана возможность построения доверительных границ для плотности вероятности с учётом результатов исследования асимптотических свойств её непараметрической оценки типа Розенблатта — Парзена [4].

Перспективное направление решения проблемы доверительного оценивания связано с использованием регрессионной оценки плотности вероятности [5, 6]. Её синтез осуществляется путём декомпозиции исходных статистических данных и анализа на основе кривой регрессии количественных характеристик получаемых множеств случайных величин [5]. Особенность структуры регрессионной оценки плотности вероятности открывает возможность построения на её основе доверительных границ для плотности вероятности и решающей функции в задаче распознавания образов.

В данной работе предлагается и исследуется алгоритмический подход доверительного оценивания байесовского уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов, основанный на использовании регрессионной оценки плотности вероятности.

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

1. Регрессионная оценка плотности вероятности. Пусть имеется выборка $V' = (x^i,\ i = \overline{1,n})$ из n независимых значений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$.

Разобьём область определения $p(x)$ на N непересекающихся интервалов длиной 2β и сформируем множества случайных величин X^j, $j = \overline{1,N}$. В качестве количественных характеристик X^j примем частоту $\tilde{P}^j$ попадания случайной величины x в j-й интервал и его центр z^j. На основе полученной информации определим массив данных $V_1 = (z^j,\ \bar{p}^j = \tilde{P}^j/(2\beta),\ j = \overline{1,N})$, составленный из центров z^j введённых интервалов и соответствующих им оценок $\bar{p}^j$ плотности вероятности. Объём N выборки V_1 может быть значительно меньше количества элементов n исходной статистической информации.

В качестве приближения по эмпирическим данным V_1 искомой плотности вероятности $p(x)$ примем статистику [5]

$$\bar{p}(x) = c^{-1} \sum_{j=1}^{N} \tilde{P}^j \Phi\Big(\frac{x - z^j}{c}\Big), \tag{1}$$

где ядерные функции $\Phi(u)$ удовлетворяют условиям

$$\Phi(u) = \Phi(-u), \quad 0 \le \Phi(u) < \infty, \quad \int\limits_{-\infty}^{+\infty} \Phi(u)du = 1, \quad \int\limits_{-\infty}^{+\infty} u^2\Phi(u)du = 1.$$

Коэффициенты размытости $c = c(N)$ ядерных функций в оценке плотности вероятности (1) убывают с ростом количества N интервалов дискретизации области определения плотности вероятности $p(x)$.

Нетрудно убедиться, что регрессионная оценка плотности $\bar{p}(x)$ является нормированной функцией, т. е. удовлетворяет основному свойству плотности вероятности, и обладает свойствами асимптотической сходимости к $p(x)$ [5]. Из условия минимума асимптотического выражения среднеквадратического отклонения $\bar{p}(x)$ от $p(x)$ получена процедура оптимального выбора количества N интервалов дискретизации [7].

2. Синтез непараметрического алгоритма распознавания образов. Рассмотрим методику построения непараметрического классификатора, соответствующего критерию максимального правдоподобия, с использованием регрессионной оценки плотности вероятности (1).

Пусть $V = V_1 \cup V_2$ — обучающая выборка, составленная из значений признака x классифицируемых объектов $V_r = (x^i,\ i = \overline{1,n_r})$, принадлежащих к одному из двух классов $\Omega_r = \Omega_r(x)$, $r = 1, 2$. Вид условных плотностей вероятностей $p_r(x)$ распределения значений x в классах Ω_r, $r = 1, 2$, неизвестен.

В данных условиях непараметрическое решающее правило распознавания образов, соответствующее критерию максимального правдоподобия, имеет вид

$$\bar{m}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) < 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) \ge 0, \end{cases} \tag{2}$$

где $\bar{f}_{12}(x) = \bar{p}_2(x) - \bar{p}_1(x)$ — непараметрическая оценка байесовского уравнения разделяющей поверхности $f_{12}(x) = p_2(x) - p_1(x)$ между классами Ω_1, Ω_2.

В качестве оценки условной плотности вероятности $p_r(x)$ будем использовать регрессионную оценку типа (1). Для этого воспользуемся методикой декомпозиции области определения $p_r(x)$, представленной в разд. 1. На этой основе преобразуем исходные данные

V_r в выборку $\bar{V}_r = (z^i,\ \tilde{P}_r^i/(2\beta_r),\ i = \overline{1, N_r})$. Здесь z^i — центр i-го интервала дискретизации области определения $p_r(x)$ длиной $2\beta_r$; N_r — количество интервалов; $\tilde{P}_r^i$ — частота попадания случайной величины x в i-й интервал.

Тогда непараметрическая оценка $\bar{f}_{12}(x)$, восстанавливаемая по выборкам $\bar{V}_r$, $r = 1,\ 2$, запишется в виде

$$\bar{f}_{12}(x) = c^{-1} \sum_{r=1}^{2} (-1)^r \sum_{i=1}^{N_r} \tilde{P}_r^i \Phi\Big(\frac{x - z^i}{c}\Big). \tag{3}$$

Оптимизация непараметрического решающего правила (2) по коэффициенту размытости c ядерных функций осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов:

$$\bar{\rho}(c) = \frac{1}{N} \sum_{t=1}^{N} 1(\sigma(t), \bar{\sigma}(t)), \qquad 1(\sigma(t), \bar{\sigma}(t)) = \begin{cases} 0, & \text{если } \sigma(t) = \bar{\sigma}(t), \\ 1, & \text{если } \sigma(t) \neq \bar{\sigma}(t), \end{cases}$$

где $N = N_1 + N_2$; $\sigma(t)$, $\bar{\sigma}(t)$ — соответственно «указания учителя» и решение алгоритма (2) о принадлежности ситуации z^t к одному из двух классов. При формировании решения $\bar{\sigma}(t)$ ситуация z^t исключается из процесса обучения в непараметрической статистике (3).

3. Методика построения доверительных границ для $\bar{f}_{12}(x)$. Известно, что верхняя $\bar{\bar{P}}_r^i$ и нижняя $\bar{P}_r^i$ границы интервальной оценки вероятности принадлежности случайной величины x к i-му интервалу дискретизации с коэффициентом доверия γ определяются выражениями [8]

$$\bar{\bar{P}}_r^i = \tilde{P}_r^i + \frac{u_{1-\alpha/2}}{\sqrt{N_r}} \sqrt{\tilde{P}_r^i (1 - \tilde{P}_r^i)}, \tag{4}$$

$$\bar{P}_r^i = \tilde{P}_r^i - \frac{u_{1-\alpha/2}}{\sqrt{N_r}} \sqrt{\tilde{P}_r^i (1 - \tilde{P}_r^i)}, \quad r = 1,\ 2, \tag{5}$$

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

Организуем вычислительный эксперимент и сформируем по значениям $(z^i,\ \tilde{P}_r^i, i = \overline{1, N_r})$ в соответствии с выражениями (4), (5) массивы данных $\bar{\bar{V}}_r = (z^i,\ \bar{\bar{P}}_r^i,\ i = \overline{1, N_r})$, $\bar{V}_r = (z^i,\ \bar{P}_r^i,\ i = \overline{1, N_r})$, $r = 1,\ 2$. По полученной информации $\bar{\bar{V}}_r$, $\bar{V}_r$, $r = 1,\ 2$, построим верхние и нижние границы

$$\bar{\bar{p}}_r(x) = c^{-1} \sum_{i=1}^{N_r} \bar{\bar{P}}_r^i \Phi\Big(\frac{x - z^i}{c}\Big), \qquad \bar{p}_r(x) = c^{-1} \sum_{i=1}^{N_r} \bar{P}_r^i \Phi\Big(\frac{x - z^i}{c}\Big)$$

для плотностей вероятности $p_r(x)$, $r = 1,\ 2$.

Для одномерного случая граница между классами Ω_1, Ω_2 определяется значением λ на оси x, которое соответствует условию $f_{12}(x) = 0$, т. е. $p_1(x) = p_2(x)$. Нетрудно заметить, что для доверительных границ $\bar{\bar{\lambda}}$, $\bar{\lambda}$ на оси значений x выполняются условия $\bar{\bar{p}}_1(x) = \bar{p}_2(x)$, $\bar{p}_1(x) = \bar{\bar{p}}_2(x)$. Причём принадлежность границы λ между классами к доверительному интервалу $D = (\bar{\lambda}, \bar{\bar{\lambda}})$ определяется правилом

$$\lambda \in D, \text{ если } \bar{\bar{p}}_2(x) > \bar{p}_1(x) \text{ и } \bar{\bar{p}}_1(x) > \bar{p}_2(x). \tag{6}$$

4. Анализ эффективности методики построения доверительных границ. Исследуем влияние объёма n обучающей выборки V и параметров процедуры её декомпозиции на эффективность методики построения доверительных границ для уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов. Условные плотности вероятности распределения значений признака x в классах Ω_1, Ω_2 определяются нормальными законами

$$p_r(x) = 1/\sqrt{2\pi}\exp(-(x-m_r)^2/2),$$

где математические ожидания случайной величины x задаются значениями $m_r = (-1)^r 1{,}5$, $r = 1,\ 2$.

Для выбора количества интервалов дискретизации области изменения значений случайной величины будем использовать формулы Хайнкольда — Гаеде [9], Брукса — Каррузера [10], Старджесса [11] соответственно:

$$N_r = \sqrt{n_r}, \tag{7}$$

$$N_r = 5\lg n_r, \tag{8}$$

$$N_r = \log_2 n_r + 1, \quad r = 1,\ 2. \tag{9}$$

Синтез регрессионной оценки плотности вероятности осуществлялся на основе ядерных функций В. А. Епанечникова [12]

$$\Phi(u) = \begin{cases} \dfrac{3}{4\sqrt{5}}\,(1-u^2/5) & \forall |u| < \sqrt{5}, \\ 0 & \forall |u| \geq \sqrt{5}. \end{cases}$$

При формировании массивов данных $\bar{\bar{V}}_r$, $\bar{V}_r$ доверительные интервалы для вероятностей P_r^j, $j = \overline{1, N_r}$, $r = 1,\ 2$, определялись с коэффициентами доверия $\gamma = 0{,}90$ и $\gamma = 0{,}95$.

При одних и тех же объёмах $n = n_1 + n_2$, $n_1 = n_2$, обучающей выборки V в соответствии с правилом (6) многократно ($m = 100$) строились доверительные границы для уравнения разделяющей поверхности. В каждом вычислительном эксперименте устанавливался факт принадлежности λ доверительному интервалу D. По полученной информации оценивалась вероятность $\bar{\gamma}_{12}$ события $\lambda \in D$. Значения $\bar{\gamma}_{12}$ можно определить как оценку коэффициента доверия при построении доверительного интервала для границы λ между классами.

Примеры доверительных границ уравнения разделяющей поверхности для конкретных условий вычислительного эксперимента приведены на рис. 1.

С увеличением n исходных данных наблюдается уменьшение длины $d = \bar{\bar{\lambda}} - \bar{\lambda}$ доверительного интервала D для границы между классами, что характерно для всех приведённых методов дискретизации (рис. 2). Этот факт объясняется увеличением в соответствии с формулами (7)–(9) объёма N обучающей выборки $\bar{V}$, используемой при синтезе непараметрической оценки уравнения разделяющей поверхности (3) и повышением её аппроксимационных свойств. При $n \leq 100$ длина доверительных интервалов рассматриваемых методов дискретизации сопоставима, так как количество N интервалов дискретизации различается незначительно. Например, при $n = 100$ значения N, вычисленные по формулам (7)–(9), равны 10, 10, 8. С увеличением n длина доверительного интервала, соответствующая методу дискретизации (7), больше, чем при использовании формул (8), (9). В рассматриваемых

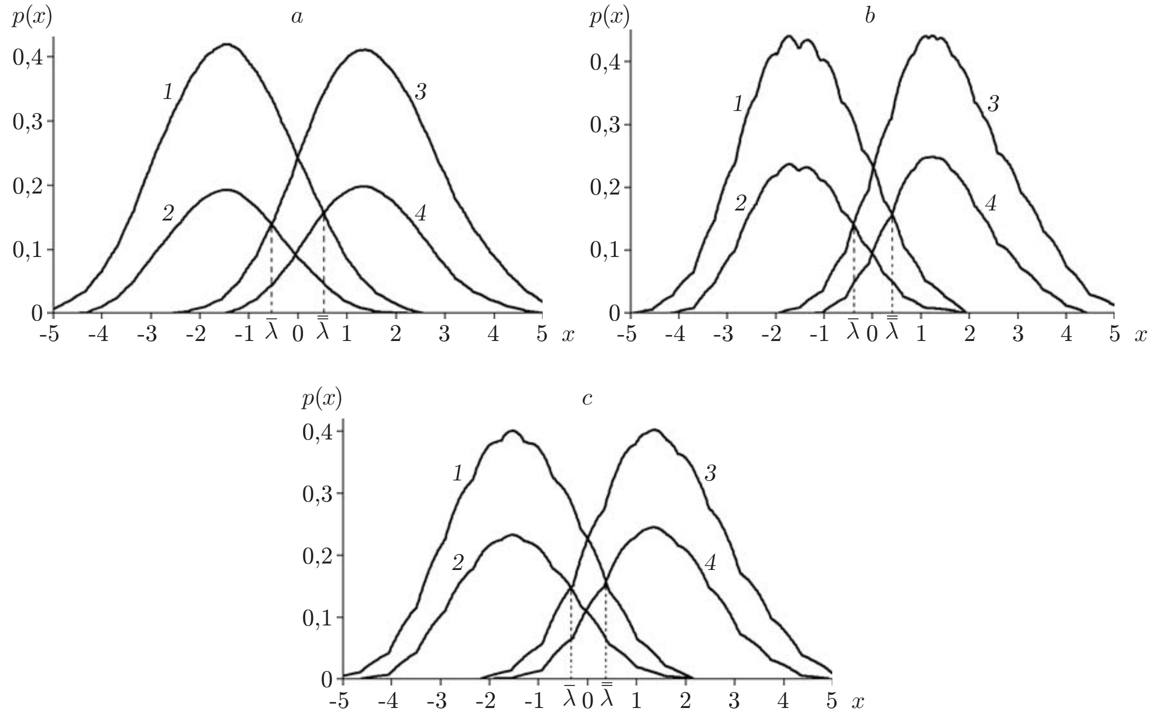

Рис. 1. Доверительные границы $\bar{\lambda}$, $\bar{\bar{\lambda}}$ для уравнения разделяющей поверхности $\lambda = f_{12}(x)$ при использовании различных методов дискретизации интервала изменения случайной величины: a — (7), b — (8), c — (9). Условия вычислительного эксперимента: $n = 500$, $\gamma = 0{,}95$. Кривые *1*, *3* и *2*, *4* соответствуют верхним $\bar{\bar{p}}_1(x)$, $\bar{\bar{p}}_2(x)$ и нижним $\bar{p}_1(x)$, $\bar{p}_2(x)$ доверительным границам плотностей вероятности $p_1(x)$, $p_2(x)$

условиях количество интервалов дискретизации N, определяемое выражением (7), значительно больше значений N, вычисляемых по формулам (8) и (9). Однако увеличение N не позволяет компенсировать ухудшение характеристик оценок вероятностей попадания случайной величины в интервалы дискретизации, что снижает аппроксимационные свойства статистики (3) и сказывается на увеличении её доверительного интервала. Данный вывод подтверждается также в процессе анализа зависимости d от n при использовании методов дискретизации (8), (9), для которых справедливо соотношение $N\,(8) > N\,(9)$.

Независимо от методов дискретизации оценка коэффициента доверия $\bar{\gamma}_{12}$ при построении доверительных границ для уравнения разделяющей поверхности близка к единице.

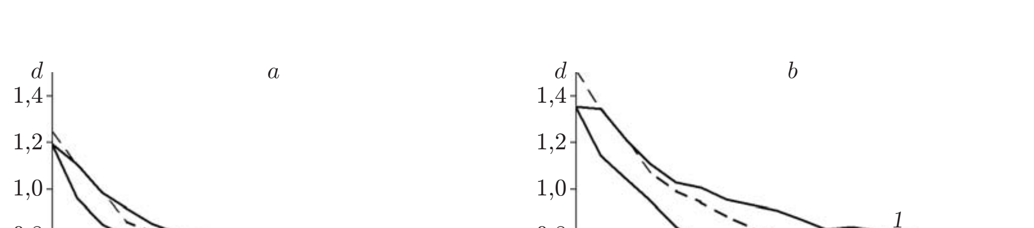

Рис. 2. Зависимость длины d доверительного интервала D для уравнения разделяющей поверхности $\lambda = f_{12}(x)$ от объёма n обучающей выборки и коэффициента доверия: $\gamma = 0{,}90$ (a), $\gamma = 0{,}95$ (b). Кривые *1–3* соответствуют методам дискретизации (7)–(9)

Заключение. Структура непараметрической оценки уравнения разделяющей поверхности, при синтезе которой используется регрессионная оценка плотности вероятности, позволяет построить доверительные границы для решающей функции в двуальтернативной задаче распознавания образов. Такой подход предполагает разбиение области значений случайной величины x на непересекающиеся интервалы в каждом классе и последующее доверительное оценивание вероятностей принадлежности x к данным интервалам по исходной статистической информации. На этой основе осуществляется синтез доверительных границ плотностей вероятности и уравнения разделяющей поверхности.

Размеры области, определяемые доверительными границами, зависят от объёма обучающей выборки, количества N интервалов дискретизации и заданного коэффициента доверия для вероятностей попадания в них случайной величины. При относительно малых n длины доверительных интервалов, соответствующие рассматриваемым методам дискретизации, сопоставимы. С увеличением n методу дискретизации с бóльшим значением N соответствует бóльшая длина доверительного интервала для уравнения разделяющей поверхности.

Предложенный подход даёт возможность обобщить полученные результаты на построение доверительных границ многомерного уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов.

СПИСОК ЛИТЕРАТУРЫ

1. **Пугачев В. С.** Теория вероятностей и математической статистики. М.: Наука, 1979. 496 с.
2. **Pearson K.** On lines and planes of closest fit to systems of points in space // Philosophical Magazine. 1901. **2**, N 6. P. 559–572.
3. **Мания Г. М.** Статистическое оценивание распределения вероятностей. Тбилиси: ТГУ, 1974. 238 с.
4. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Stat. 1962. **33**, N 3. P. 1065–1076.
5. **Лапко А. В., Лапко В. А.** Регрессионная оценка многомерной плотности вероятности и её свойства // Автометрия. 2014. **50**, № 2. С. 50–56.
6. **Lapko A. V., Lapko V. A.** Construction of confidence limits for the probability density function on the basis of nonparametric estimation of the function // Measur. Techn. 2014. **56**, N 12. P. 1354–1357.
7. **Lapko A. V., Lapko V. A.** Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density // Measur. Techn. 2013. **56**, N 7. P. 763–767.
8. **Математическая** статистика: Учебник для вузов /Под ред. В. С. Зарубина, А. П. Крищенко. М.: МГТУ им. Н. Э. Баумана, 2001. 424 с.
9. **Heinhold J., Gaede K.-W.** Ingenieur-Statistic. München — Wien: Oldenbourg, 1964. 352 p.
10. **Шторм Р.** Теория вероятностей. Математическая статистика. Статистический контроль качества. М.: Мир, 1970. 368 с.
11. **Sturges H. A.** The choice of a class interval // Journ. Amer. Stat. Association. 1926. **21**, Is. 153. P. 65–66.
12. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, № 1. С. 156–161.

Поступила в редакцию 23 мая 2014 г.