УДК 681.513

СВОЙСТВА НЕПАРАМЕТРИЧЕСКОЙ РЕШАЮЩЕЙ ФУНКЦИИ ПРИ НАЛИЧИИ АПРИОРНЫХ СВЕДЕНИЙ О НЕЗАВИСИМОСТИ ПРИЗНАКОВ КЛАССИФИЦИРУЕМЫХ ОБЪЕКТОВ

А. В. Лапко[1], **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. Решетнева,
660014, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Исследуются асимптотические свойства непараметрической оценки уравнения разделяющей поверхности, определённой в пространстве независимых признаков классифицируемых объектов. На этой основе устанавливается значимость влияния априорных сведений о независимости случайных величин на аппроксимационные свойства непараметрической решающей функции в двухальтернативной задаче распознавания образов.

Ключевые слова: распознавание образов, независимые случайные величины, непараметрическая оценка, асимптотические свойства, априорная информация.

Введение. Непараметрические алгоритмы распознавания образов, основанные на оценках плотности вероятности типа Розенблатта — Парзена, широко используются при исследовании объектов различной природы в условиях априорной неопределённости [1].

Разработан ряд модификаций непараметрических классификаторов, ориентированных на условия малых [2] и больших [3, 4] объёмов обучающих выборок и их неоднородный характер [5].

Дальнейшее развитие непараметрических методов распознавания образов заключается в наиболее полном учёте априорных сведений об особенностях изучаемых закономерностей и информации, содержащейся в обучающих выборках [6, 7].

Цель данной работы состоит в обосновании возможности повышения аппроксимационных свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов за счёт использования априорной информации о независимости признаков классифицируемых объектов.

Непараметрическая оценка уравнения разделяющей поверхности. Для получения аналитически значимых результатов без существенной потери общности рассмотрим методику построения непараметрического классификатора на примере двухальтернативной задачи распознавания образов в пространстве двух независимых признаков $x = (x_1, x_2)$.

Пусть $V = (x^i, \sigma(i),\ i = \overline{1, n}\,)$ — обучающая выборка объёма n, составленная из признаков $x^i = (x_1^i, x_2^i)$ классифицируемых объектов и соответствующих им «указаний учителя» $\sigma(i)$ об их принадлежности к одному из двух классов Ω_1, Ω_2. Вид условных плотностей вероятности $p_r(x_v)$ распределения признаков x_v, $v = 1, 2$, анализируемых объектов в классах Ω_r, $r = 1, 2$, априори неизвестен.

В данных условиях непараметрическое решающее правило распознавания образов,

соответствующее критерию максимального правдоподобия, имеет вид

$$\bar{m}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \le 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0, \end{cases} \tag{1}$$

где

$$\bar{f}_{12}(x) = \bar{p}_2(x_1)\bar{p}_2(x_2) - \bar{p}_1(x_1)\bar{p}_1(x_2) \tag{2}$$

— непараметрическая оценка байесовского уравнения разделяющей поверхности

$$f_{12}(x) = p_2(x_1)p_2(x_2) - p_1(x_1)p_1(x_2)$$

между классами Ω_1, Ω_2.

Для оценивания плотности вероятности $p_r(x_v)$ по данным обучающей выборки V используем статистику типа Розенблатта — Парзена для одномерного случая [8]

$$p_r(x_v) = \frac{1}{n_r c} \sum_{i \in I_r} \Phi\Big(\frac{x_v - x_v^i}{c}\Big), \quad v = 1,\, 2, \quad r = 1,\, 2, \tag{3}$$

где $n_r = |I_r|$ — количество элементов множества номеров I_r ситуаций из обучающей выборки, принадлежащих к классу Ω_r.

Ядерные функции $\Phi(u)$ в непараметрической оценке плотности вероятности (3), удовлетворяющие условиям H, имеют вид

$$\Phi(u) = \Phi(-u), \qquad 0 \le \Phi(u) < \infty,$$

$$\int \Phi(u)du = 1, \qquad \int u^2\Phi(u)du = 1,$$

$$\int u^m\Phi(u)du < \infty, \quad 0 \le m < \infty.$$

Здесь и далее бесконечные пределы интегрирования опускаются.

Значения коэффициентов размытости $c = c(n)$ ядерных функций в непараметрической оценке $\bar{f}_{12}(x)$ уравнения разделяющей поверхности убывают с ростом объёма n обучающей выборки. Для упрощения последующего анализа аппроксимационных свойств $\bar{f}_{12}(x)$ считается, что интервалы изменения признаков x_1, x_2 классифицируемых объектов сопоставимы. Поэтому появляется возможность полагать значения коэффициентов размытости $c_v(n)$, $v = 1,\, 2$, соответствующих признакам x_1, x_2, равными: $c_1(n) = c_2(n) = c(n)$.

Оптимизация непараметрического решающего правила (1) по коэффициенту размытости ядерных функций c осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов

$$\bar{\rho}(c) = \frac{1}{n}\sum_{t=1}^{n} 1(\sigma(t),\, \bar{\sigma}(t)),$$

$$1(\sigma(t),\, \bar{\sigma}(t)) = \begin{cases} 0, & \text{если } \sigma(t) = \bar{\sigma}(t), \\ 1, & \text{если } \sigma(t) \ne \bar{\sigma}(t), \end{cases}$$

где $\bar{\sigma}(t)$ — «решение» о принадлежности ситуации x^t к одному из двух классов, полученное с помощью алгоритма (1). При формировании решения $\bar{\sigma}(t)$ ситуация x^t исключается из процесса обучения в непараметрической статистике (2).

Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности. Справедливо следующее утверждение.

Теорема. Пусть плотности вероятностей $p_r(x_v)$, $r = 1, 2$, $v = 1, 2$, распределения независимых признаков x_1, x_2 анализируемых объектов в классах Ω_1, Ω_2 и первые две их производные ограничены и непрерывны; ядерные функции $\Phi(u)$ удовлетворяют условиям нормированности, положительности и симметричности H; последовательность $c(n) = c$ коэффициентов размытости ядерных функций такова, что при $n_1 \to \infty$, $n_2 \to \infty$ значения $c \to 0$, а $n_1 c \to \infty$ и $n_2 c \to \infty$. Тогда непараметрическая оценка $\bar{f}_{12}(x)$ байесовского уравнения разделяющей поверхности $f_{12}(x) = p_2(x_1)p_2(x_2) - p_1(x_1)p_1(x_2)$ обладает свойством асимптотической несмещённости и состоятельности.

Доказательство.

1. По определению имеем

$$\mathrm{M}(\bar{f}_{12}(x)) = \mathrm{M}(\bar{p}_2(x_1)\bar{p}_2(x_2) - \bar{p}_1(x_1)\bar{p}_1(x_2)) =$$

$$= \frac{1}{n_2^2 c^2} \sum_{i \in I_2} \sum_{j \in I_2} \mathrm{M}\Big(\Phi\Big(\frac{x_1 - x_1^i}{c}\Big)\Phi\Big(\frac{x_2 - x_2^j}{c}\Big)\Big) -$$

$$- \frac{1}{n_1^2 c^2} \sum_{i \in I_1} \sum_{j \in I_1} \mathrm{M}\Big(\Phi\Big(\frac{x_1 - x_1^i}{c}\Big)\Phi\Big(\frac{x_2 - x_2^j}{c}\Big)\Big) =$$

$$= \frac{1}{n_2^2 c^2} \sum_{i \in I_2} \sum_{j \in I_2} \int\int \Phi\Big(\frac{x_1 - x_1^i}{c}\Big)\Phi\Big(\frac{x_2 - x_2^j}{c}\Big) p_2(x_1^i)p_2(x_2^j)dx_1^i dx_2^j -$$

$$- \frac{1}{n_1^2 c^2} \sum_{i \in I_1} \sum_{j \in I_1} \int\int \Phi\Big(\frac{x_1 - x_1^i}{c}\Big)\Phi\Big(\frac{x_2 - x_2^i}{c}\Big) p_1(x_1^i)p_1(x_2^j)dx_1^i dx_2^j =$$

$$= \frac{1}{c^2} \int \Phi\Big(\frac{x_1 - t_1}{c}\Big) p_2(t_1)dt_1 \int \Phi\Big(\frac{x_2 - t_2}{c}\Big) p_2(t_2)dt_2 -$$

$$- \frac{1}{c^2} \int \Phi\Big(\frac{x_1 - t_1}{c}\Big) p_1(t_1)dt_1 \int \Phi\Big(\frac{x_2 - t_2}{c}\Big) p_1(t_2)dt_2,$$

где M — знак математического ожидания.

При выполнении данных преобразований учитывается, что элементы статистических выборок, определяющих каждый r-й класс, являются значениями одной и той же случайной величины $t = (t_1, t_2)$ с плотностью вероятности $p_r(t) = p_r(t_1)p_r(t_2)$, $r = 1, 2$.

Проведём в составляющих последнего выражения замену переменных $x_1 - t_1 = u_1 c$, $x_2 - t_2 = u_2 c$ и, разлагая функции $p_r(x_1 - u_1 c)$, $p_r(x_2 - u_2 c)$, $r = 1, 2$, в ряд Тейлора в точке $x = (x_1, x_2)$, с учётом свойств ядерной функции при достаточно больших значениях n_1, n_2 получим

$$\mathrm{M}(\bar{f}_{12}(x) - f_{12}(x)) =$$

$$= \frac{c^2}{2}\Big(\sum_{r=1}^{2}(-1)^r p_r^{(2)}(x_2)p_r(x_1) + \sum_{r=1}^{2}(-1)^r p_r^{(2)}(x_1)p_r(x_2)\Big) +$$

$$+ \frac{c^4}{4}\sum_{r=1}^{2}(-1)^r p_r^{(2)}(x_1)p_r^{(2)}(x_2) + 0(c^6). \qquad (4)$$

Здесь $p_r^{(2)}(x_1)$, $p_r^{(2)}(x_2)$ — вторые производные плотностей вероятности $p_r(x_1)$, $p_r(x_2)$, $r = 1, 2$, по переменным x_1, x_2 соответственно; символом $0(c^6)$ обозначены слагаемые порядка малости c^6 и выше.

Отсюда из условия $c \to 0$ при $n_1 \to \infty$, $n_2 \to \infty$ следует свойство асимптотической несмещённости непараметрической статистики $\bar{f}_{12}(x)$.

2. Для доказательства состоятельности непараметрической оценки $\bar{f}_{12}(x)$ уравнения разделяющей поверхности между классами исследуем асимптотические свойства среднеквадратического отклонения

$$\mathrm{M}\int\int(\bar{f}_{12}(x) - f_{12}(x))^2 dx_1 dx_2 = \int\int \mathrm{M}(p_2(x_1)p_2(x_2) - \bar{p}_2(x_1)\bar{p}_2(x_2))^2 dx_1 dx_2 -$$

$$- 2\int\int \mathrm{M}((p_2(x_1)p_2(x_2) - \bar{p}_2(x_1)\bar{p}_2(x_2))(p_1(x_1)p_1(x_2) - \bar{p}_1(x_1)\bar{p}_1(x_2)))dx_1 dx_2 +$$

$$+ \int\int \mathrm{M}(p_1(x_1)p_1(x_2) - \bar{p}_1(x_1)\bar{p}_1(x_2))^2 dx_1 dx_2. \qquad (5)$$

Асимптотическое выражение для среднеквадратического отклонения непараметрической оценки $\bar{p}_r(x_1)\bar{p}_r(x_2)$ произведения плотностей вероятности $p_r(x_1)p_r(x_2)$, $r = 1, 2$, получено в работе [7]:

$$\int\int \mathrm{M}(p_r(x_1)p_r(x_2) - \bar{p}_r(x_1)\bar{p}_r(x_2))^2 dx_1 dx_2 \sim$$

$$\sim \frac{\|\Phi(u)\|^2\|p_r(x_2)\|^2}{n_r c} + \frac{\|\Phi(u)\|^2\|p_r(x_1)\|^2}{n_r c} +$$

$$+ \frac{c^4}{4}\int\int(p_r(x_2)p_r^{(2)}(x_1) + p_r(x_1)p_r^{(2)}(x_2))^2 dx_1 dx_2, \qquad (6)$$

где

$$\|\Phi(u)\|^2 = \int \Phi^2(u)du; \quad \|p_r(x_v)\|^2 = \int p_r^2(x_v)dx_v, \quad v = 1, 2, \ r = 1, 2.$$

При формировании выражения (6) величины порядка малости $0(1/n_r)$, $0(1/n_r^2c^2)$, $0(c/n_r)$, $r = 1, 2$; $0(c^6)$ не учитываются.

Выполняя последовательный анализ составляющих второго слагаемого среднеквадратического отклонения (5), получим его асимптотическое выражение. С учётом независимости случайных величин x_1, x_2 и результатов исследований [9] в условиях $k = 1$ при достаточно больших значениях n_1, n_2 запишем

$$\mathrm{M}(p_1(x_1)p_1(x_2)\bar{p}_2(x_1)\bar{p}_2(x_2)) \sim$$

$$\sim p_1(x_1)p_1(x_2)\Big(p_2(x_1) + \frac{c^2}{2}p_2^{(2)}(x_1)\Big)\Big(p_2(x_2) + \frac{c^2}{2}p_2^{(2)}(x_2)\Big), \tag{7}$$

$$\mathrm{M}(p_2(x_1)p_2(x_2)\bar{p}_1(x_1)\bar{p}_1(x_2)) \sim$$

$$\sim p_2(x_1)p_2(x_2)\Big(p_1(x_1) + \frac{c^2}{2}p_1^{(2)}(x_1)\Big)\Big(p_1(x_2) + \frac{c^2}{2}p_1^{(2)}(x_2)\Big). \tag{8}$$

Следуя технологии преобразований, использованной при доказательстве первой части теоремы, нетрудно показать, что при $n_1 \to \infty$ и $n_2 \to \infty$

$$\mathrm{M}(\bar{p}_1(x_1)\bar{p}_1(x_2)\bar{p}_2(x_1)\bar{p}_2(x_2)) = \mathrm{M}(\bar{p}_1(x_1)\bar{p}_2(x_1))\mathrm{M}(\bar{p}_1(x_2)\bar{p}_2(x_2)) \sim$$

$$\sim \Big(p_1(x_1) + \frac{c^2}{2}p_1^{(2)}(x_1)\Big)\Big(p_2(x_1) + \frac{c^2}{2}p_2^{(2)}(x_1)\Big)\Big(p_1(x_2) + \frac{c^2}{2}p_1^{(2)}(x_2)\Big)\Big(p_2(x_2) + \frac{c^2}{2}p_2^{(2)}(x_2)\Big). \tag{9}$$

Подставляя (7)–(9) во второе слагаемое (5), получим его асимптотическое выражение

$$-\frac{c^4}{2}\int\int (p_1(x_1)p_1^{(2)}(x_2) + p_1(x_2)p_1^{(2)}(x_1))(p_2(x_1)p_2^{(2)}(x_2) + p_2(x_2)p_2^{(2)}(x_1))dx_1dx_2.$$

На этой основе, принимая во внимание результаты (6), окончательно запишем асимптотическое выражение среднеквадратического отклонения $\bar{f}_{12}(x)$ от $f_{12}(x)$:

$$\mathrm{M}\int\int (\bar{f}_{12}(x) - f_{12}(x))^2 dx_1dx_2 \sim$$

$$\sim \frac{\|\Phi(u)\|^2}{c}\Big(\frac{\|p_1(x_1)\|^2 + \|p_1(x_2)\|^2}{n_1} + \frac{\|p_2(x_1)\|^2 + \|p_2(x_2)\|^2}{n_2}\Big) + \frac{c^4}{4}\bar{B}, \tag{10}$$

где

$$\bar{B} = \int\int [(p_1(x_1)p_1^{(2)}(x_2) + p_1(x_2)p_1^{(2)}(x_1)) - (p_2(x_1)p_2^{(2)}(x_2) + p_2(x_2)p_2^{(2)}(x_1))]^2 dx_1dx_2.$$

Заметим, что при выполнении условий $c \to 0$, $n_1c \to \infty$ и $n_2c \to \infty$ для $n_1 \to \infty$ и $n_2 \to \infty$ непараметрическая оценка $\bar{f}_{12}(x)$ сходится в среднеквадратическом к байесовскому уравнению разделяющей поверхности $f_{12}(x)$, а с учётом свойства её асимптотической несмещённости является состоятельной.

Анализ свойств статистики $\bar{f}_{12}(x)$. Исследуем влияние априорных сведений о независимости признаков классифицируемых объектов на аппроксимационные свойства непараметрической оценки уравнения разделяющей поверхности. Для этого при оптимальных значениях коэффициентов размытости ядерных функций сравним условия асимптотической сходимости в среднеквадратическом статистики (2) и традиционной непараметрической оценки уравнения разделяющей поверхности

$$\tilde{f}_{12}(x) = \tilde{p}_2(x_1, x_2) - \bar{p}_1(x_1, x_2),$$

где

$$\tilde{p}_r(x_1, x_2) = \frac{1}{n_r c^2} \sum_{i \in I_r} \prod_{v=1}^{2} \Phi\Big(\frac{x_v - x_v^i}{c}\Big), \quad r = 1, 2.$$

В работе [10] установлено, что максимальная эффективность непараметрической решающей функции в задаче распознавания образов достигается при равномерном распределении элементов обучающей выборки между классами. Поэтому определим оптимальные значения $\bar{c}$ статистики (2) из условия минимума выражения (10) при $n_1 = n_2 = n/2$. После очевидных преобразований получим

$$\bar{c} = \Big(\frac{2\|\Phi(u)\|^2 \bar{A}}{n\bar{B}}\Big)^{1/5}, \tag{11}$$

где

$$\bar{A} = \sum_{r=1}^{2} (\|p_r(x_1)\|^2 + \|p_r(x_2)\|^2).$$

Подставляя значения $\bar{c}$ в выражение (10), вычислим его минимальное значение

$$\bar{W} = \frac{5}{4}\Big[\Big(\frac{2\|\Phi(u)\|^2 \bar{A}}{n}\Big)^4 \bar{B}\Big]^{1/5}.$$

В соответствии с результатами исследований [11] минимальное среднеквадратическое отклонение традиционной непараметрической оценки $\tilde{f}_{12}(x)$ при $k = 2$ и $n_1 = n_2$ определяется выражением

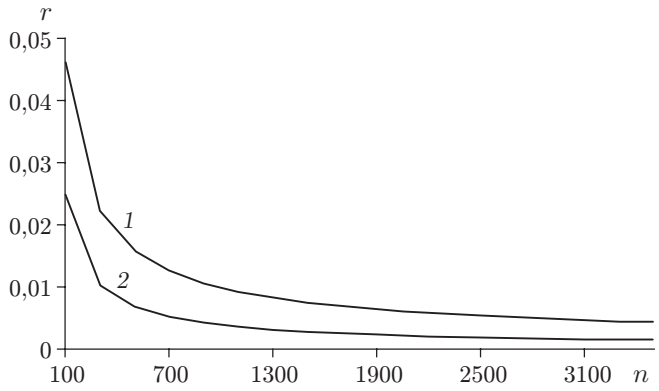
$$\tilde{W} = 3\Big[\Big(\frac{(\|\Phi(u)\|^2)^2}{n}\Big)^4 \tilde{B}^2\Big]^{1/6},$$

где

$$\tilde{B} = \int\int \Big(\sum_{v=1}^{2} (p_{2v}^{(2)}(x_1, x_2) - p_{1v}^{(2)}(x_1, x_2))\Big)^2 dx_1 dx_2;$$

$p_{rv}^{(2)}(x_1, x_2)$ — вторая производная плотности вероятности $p_r(x_1, x_2)$ по переменной x_v, $v = 1, 2$.

При конечных $p_r(x_v), p_r^{(2)}(x_v)$, $v = 1, 2$, $r = 1, 2$, с ростом объёма n обучающей выборки значения $\bar{W}$ минимального среднеквадратического отклонения $\bar{f}(x_1, x_2)$ стремятся к нулю пропорционально величине $r = n^{-4/5}$. Причём порядок подобной сходимости выше, чем



Зависимость порядка сходимости r значений $\tilde{W}$ (кривая *1*) и $\bar{W}$ (кривая *2*) к нулю от объёма n статистических данных при синтезе непараметрических оценок $\tilde{f}(x_1, x_2)$, $\bar{f}(x_1, x_2)$ уравнений разделяющей поверхности в двухальтернативной задаче распознавания образов

для $\tilde{W}$, значения которого убывают пропорционально величине $n^{-4/6}$ (см. рисунок). С учётом взаимосвязи между среднеквадратическим отклонением и дисперсией нетрудно заметить, что данная закономерность свойственна и для дисперсий непараметрических оценок решающих функций $\bar{f}(x_1, x_2)$, $\tilde{f}(x_1, x_2)$.

Полученные результаты аналитических исследований обосновывают значимость информации о независимости случайных величин при синтезе непараметрических решающих правил классификации. Кроме того, открывается возможность количественной оценки этой значимости.

Заключение. Наличие априорных сведений о независимости признаков классифицируемых объектов позволяет повысить аппроксимационные свойства непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов. Подобная информация может быть определена на этапе предварительного анализа обучающей выборки путём постановки и проверки статистических гипотез о независимости случайных величин. Полученные результаты являются основой для синтеза эффективных непараметрических алгоритмов классификации, соответствующих критерию максимального правдоподобия.

СПИСОК ЛИТЕРАТУРЫ

1. **Лапко А. В., Лапко В. А.** Непараметрические системы обработки неоднородной информации. Новосибирск: Наука, 2007. 197 с.
2. **Лапко А. В., Лапко В. А., Шарков М. А.** Непараметрические методы обнаружения закономерностей в условиях малых выборок // Изв. вузов. Сер. Приборостроение. 2008. **51**, № 8. С. 62–67.
3. **Лапко А. В., Лапко В. А.** Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. 2010. **46**, № 1. С. 70–78.
4. **Лапко А. В., Лапко В. А.** Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. 2011. **47**, № 4. С. 76–82.
5. **Лапко А. В., Лапко В. А.** Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. 2008. **44**, № 3. С. 65–74.
6. **Lapko A. V., Lapko V. A.** Hybrid systems of pattern recognition // Pattern Recogn. and Image Analysis. 2008. **18**, N 1. P. 7–13.

7. **Лапко А. В., Лапко В. А.** Непараметрическая оценка плотности вероятности независимых случайных величин // Информатика и системы управления. 2011. **29**, № 3. С. 118–124.
8. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. **33**, N 3. P. 1065–1076.
9. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, № 1. С. 156–161.
10. **Лапко А. В., Лапко В. А.** Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов // Автометрия. 2010. **46**, № 3. С. 48–53.
11. **Лапко А. В., Лапко В. А.** Асимптотические свойства многомерной непараметрической оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. 2010. **39**, № 1. С. 16–19.

Поступила в редакцию 16 февраля 2012 г.