УДК 681.513

АНАЛИЗ АСИМПТОТИЧЕСКИХ СВОЙСТВ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ УРАВНЕНИЯ РАЗДЕЛЯЮЩЕЙ ПОВЕРХНОСТИ В ДВУХАЛЬТЕРНАТИВНОЙ ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ*

А. В. Лапко, В. А. Лапко

Учреждение Российской академии наук
Институт вычислительного моделирования Сибирского отделения РАН,
630036, г. Красноярск, Академгородок, 50, стр. 44
E-mail: lapko@icm.krasn.ru

Исследуются количественные зависимости аппроксимационных свойств непараметрических оценок уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов от неравномерности распределения элементов обучающей выборки между классами.

Ключевые слова: непараметрическая статистика, распознавание образов, асимптотические свойства, обучающая выборка, решающая функция.

Введение. Применение теории классификации и методов непараметрической статистики — одно из перспективных направлений исследования систем в условиях априорной неопределённости. Его значимость состоит в возможности создания универсальных математических средств, адаптируемых к условиям исследования систем различной природы [1, 2].

Дальнейшее развитие непараметрических методов распознавания образов делает необходимым углублённое изучение их свойств, позволяющих установить количественную связь между показателями эффективности алгоритмов и условиями классификации.

В данной работе на основе анализа свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов определяются количественные зависимости её аппроксимационных свойств от степени неравномерности распределения элементов обучающей выборки между классами.

Синтез непараметрического алгоритма распознавания образов. Рассмотрим методику построения непараметрического классификатора на примере двухальтернативной задачи распознавания образов в пространстве непрерывного признака x.

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

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

где

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

*Работа выполнена при поддержке Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг. (ГК № 02.740.11.0621).

— уравнение 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 относительно вида законов распределения $p_j(x)$, $j = 1, 2$, при статистическом оценивании уравнения разделяющей поверхности используются непараметрические методы статистики.

Пусть $V = (x^i, \sigma(i), i = \overline{1, n})$ — обучающая выборка объёма n, составленная из признака x^i классифицируемых объектов и соответствующих им «указаний учителя» $\sigma(i)$ об их принадлежности к одному из двух классов Ω_1, Ω_2.

В качестве оценки условной плотности вероятности $p_j(x)$ по данным обучающей выборки V используем статистику типа Розенблатта — Парзена для одномерного случая [4]:

$$p_j(x) = \frac{1}{n_j c} \sum_{i \in I_j} \Phi\Big(\frac{x - x^i}{c}\Big), \quad j = 1, 2, \tag{3}$$

где $n_j = |I_j|$ — количество элементов множества номеров I_j ситуаций из обучающей выборки, принадлежащих к классу Ω_j; $\Phi(u)$ — ядерные функции, удовлетворяющие условиям H:

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

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

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

c — коэффициент размытости ядерных функций, значения которого убывают с ростом количества n_j элементов множества I_j, $j = 1, 2$. Здесь и далее бесконечные пределы интегрирования опускаются.

Тогда непараметрическая оценка уравнения разделяющей поверхности $f_{12}(x)$ (2) представляется в виде [1, 2]

$$\bar{f}_{12}(x) = [nc]^{-1} \sum_{i=1}^{n} \sigma_1(i)\Phi\Big(\frac{x - x^i}{c}\Big), \tag{4}$$

где

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

$\bar{P}_j = n_j/n$ — оценка априорной вероятности принадлежности ситуаций обучающей выборки к классу Ω_j, $j = 1, 2$.

Оптимизация непараметрического решающего правила

$$\bar{m}(x): \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \le 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0, \end{cases} \tag{5}$$

по коэффициенту размытости ядерных функций c осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов

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

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

где $\bar{\sigma}(t)$ — решение о принадлежности ситуации x^t к одному из двух классов, принятое с помощью алгоритма (5). При формировании решения $\bar{\sigma}(t)$ ситуация x^t исключается из процесса обучения в непараметрической статистике (4).

Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности. Асимптотические свойства статистики (4) 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 H; последовательность $c(n) = c$ коэффициентов размытости ядерных функций такова, что при $n_1 \to \infty$, $n_2 \to \infty$ значения $c \to 0$ и $\dfrac{n_1 + n_2}{n_1 n_2 c} \to 0$. Тогда непараметрическая оценка $\bar{f}_{12}(x)$ уравнения разделяющей поверхности $f_{12}(x)$ обладает свойством асимптотической несмещённости и состоятельности.

Доказательство.

1. По определению имеем

$$\mathrm{M}(\bar{f}_{12}(x)) = \mathrm{M}(\bar{p}_2(x) - \bar{p}_1(x)) = \frac{1}{c}\int \Phi\Big(\frac{x-t}{c}\Big)p_2(t)dt - \frac{1}{c}\int \Phi\Big(\frac{x-t}{c}\Big)p_1(t)dt,$$

где M — знак математического ожидания.

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

$$W_1 = \mathrm{M}(\bar{f}_{12}(x) - f_{12}(x)) \sim \frac{c^2}{2}(p_2^{(2)}(x) - p_1^{(2)}(x)), \tag{6}$$

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

Из условия $c \to 0$ в выражении (6) при $n_1 \to \infty$, $n_2 \to \infty$ следует свойство асимптотической несмещённости непараметрической статистики $\bar{f}_{12}(x)$ (4).

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

$$\mathrm{M}\int(\bar{f}_{12}(x) - f_{12}(x))^2 dx = \mathrm{M}\int(p_1(x) - \bar{p}_1(x))^2 dx -$$

$$-2\mathrm{M}\int(p_1(x) - \bar{p}_1(x))(p_2(x) - \bar{p}_2(x))dx + \mathrm{M}\int(p_2(x) - \bar{p}_2(x))^2 dx. \tag{7}$$

Асимптотическое выражение для среднеквадратического отклонения $\bar{p}_j(x)$ от $p_j(x)$ получено в работе [5]:

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

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

С учётом (6) найдём асимптотическое выражение для функционала

$$\mathrm{M}\int(p_1(x)-\bar{p}_1(x))(p_2(x)-\bar{p}_2(x))dx \sim \frac{c^4}{4}\int p_1^{(2)}(x)p_2^{(2)}(x)dx.$$

Тогда при достаточно больших значениях n_1, n_2 асимптотическое выражение для среднеквадратического отклонения (7) представляется в виде

$$\mathrm{M}\int(\bar{f}_{12}(x)-f_{12}(x))^2dx \sim \frac{\|\Phi(u)\|^2(n_1+n_2)}{n_1n_2c}+\frac{c^4}{4}\left\|p_1^{(2)}(x)-p_2^{(2)}(x)\right\|^2. \tag{8}$$

Нетрудно заметить, что при выполнении условий $c\to 0$, $\dfrac{n_1+n_2}{n_1n_2c}\to 0$ при $n_j\to\infty$, $j=1,\,2$, непараметрическая оценка $\bar{f}_{12}(x)$ сходится в среднеквадратическом к байесовскому уравнению разделяющей поверхности (2), а с учётом свойства её асимптотической несмещённости является состоятельной оценкой.

Анализ асимптотических свойств $\bar{f}_{12}(x)$. Определим минимальное значение W_2 выражения (8) при оптимальных значениях c^* коэффициентов размытости ядерных функций непараметрической оценки уравнения разделяющей поверхности $\bar{f}_{12}(x)$. Приравнивая производную выражения (8) по c к нулю, получим

$$c^*=\Big(\frac{\|\Phi(u)\|^2(n_1+n_2)}{n_1n_2\|p_1^{(2)}(x)-p_2^{(2)}(x)\|^2}\Big)^{1/5}.$$

Тогда

$$W_2=\frac{5}{4}\Big[\Big(\frac{\|\Phi(u)\|^2(n_1+n_2)}{n_1n_2}\Big)^4\left\|p_1^{(2)}(x)-p_2^{(2)}(x)\right\|^2\Big]^{1/5}. \tag{9}$$

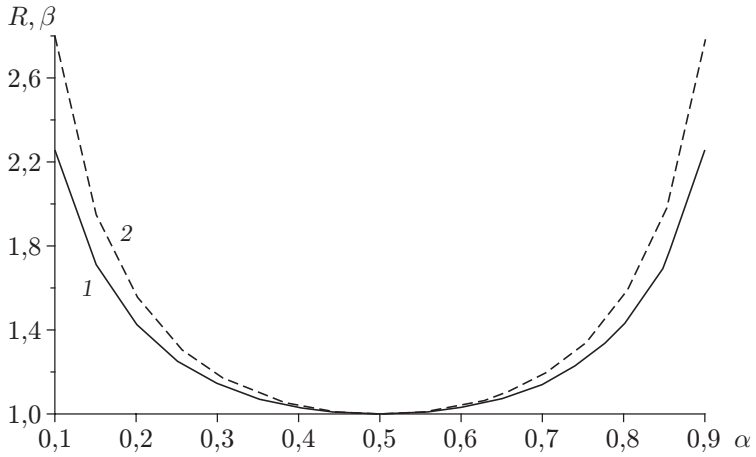
Исследуем зависимость W_2 от степени неравномерности распределения элементов обучающей выборки V между классами. Положим $n_1=\alpha n$, $n_2=(1-\alpha)n$ и перепишем выражение (9) в виде

$$W_2(\alpha,n)=\frac{5}{4}\Big[\Big(\frac{\|\Phi(u)\|^2}{\alpha(1-\alpha)n}\Big)^4\left\|p_1^{(2)}(x)-p_2^{(2)}(x)\right\|^2\Big]^{1/5}.$$

Рассмотрим отношение

$$R=\frac{W_2(\alpha,n)}{W_2(\alpha=0{,}5,\,n)}=\Big(\frac{1}{4\alpha(1-\alpha)}\Big)^{4/5}, \tag{10}$$

зависимость которого от значений α приведена на рисунке.



Зависимости отношения R (10) (кривая *1*) и коэффициента β (11) (кривая *2*) от степени α неравномерности распределения элементов обучающей выборки между классами

Минимальное значение отношения (10) достигается при $\alpha = 0{,}5$, что соответствует равным априорным вероятностям распределения элементов обучающей выборки между классами. Если такое распределение неравновероятно, то эффективность непараметрического алгоритма распознавания образов (5) снижается по сравнению с условием $n_1 = n_2$. Для компенсации данной тенденции необходимо увеличить объём обучающей выборки.

Пусть в отношении

$$\bar{R} = \frac{W_2(\alpha, n^1)}{W_2(\alpha_1 = 0{,}5,\ n)}$$

типа (10) значение $\alpha_1 \neq 0{,}5$. Определим требуемый объём n^1 обучающей выборки, при котором эффективность непараметрических алгоритмов в условиях (α, n^1) и $(\alpha_1 = 0{,}5,\ n)$ будет одинакова.

Из анализа соотношения $\bar{R} = 1$ получим

$$n^1 = n(4\alpha(1-\alpha))^{-1} = \beta n. \tag{11}$$

Зависимость коэффициента β от степени α неравномерности данных также приведена на рисунке.

Рассмотрим использование полученных результатов на следующем примере. Пусть $\alpha = 0{,}2$, при которой количество элементов первого класса обучающей выборки $n_1 = 0{,}2n$, второго — $n_2 = 0{,}8n$. В этих условиях среднеквадратическая ошибка аппроксимации байесовского уравнения разделяющей поверхности (2) (статистика (4)) в 1,43 раза больше по сравнению с равномерным распределением элементов обучающей выборки между классами. Для её снижения до уровня условий $n_1 = n_2 = 0{,}5n$ необходимо увеличить объём обучающей выборки до значения $n^1 = 1{,}56n$.

Заключение. Эффективность непараметрических алгоритмов распознавания образов зависит от степени неравномерности распределения элементов обучающей выборки между классами. Минимальное значение асимптотического выражения среднеквадратического отклонения непараметрической оценки уравнения разделяющей поверхности, соответствующей критерию максимального правдоподобия, достигается при равномерном распределении элементов обучающей выборки между классами. При нарушении данного условия повышение эффективности непараметрических решающих правил классификации

может осуществляться за счёт увеличения объёма обучающей выборки. В представленной работе получены количественные критерии оценивания эффективности непараметрических решающих функций. Дальнейшее развитие предлагаемой методики связано с её обобщением на многомерную задачу классификации и использованием решающих функций, соответствующих критерию максимума апостериорной вероятности.

СПИСОК ЛИТЕРАТУРЫ

1. **Медведев А. В.** Непараметрические системы адаптации. Новосибирск: Наука, 1983. 174 с.
2. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000. 240 с.
3. **Цыпкин Я. З.** Основы теории обучающихся систем. М.: Наука, 1970. 280 с.
4. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
5. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, вып. 1. С. 156–161.

Поступила в редакцию 9 октября 2009 г.