РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2008, том 44, № 3

УДК 681.513

АНАЛИЗ НЕПАРАМЕТРИЧЕСКИХ АЛГОРИТМОВ РАСПОЗНАВАНИЯ ОБРАЗОВ В УСЛОВИЯХ ПРОПУСКА ДАННЫХ*

А. В. Лапко, В. А. Лапко

Институт вычислительного моделирования СО РАН, г. Красноярск
E-mail: lapko@icm.krasn.ru

Исследуются непараметрические алгоритмы распознавания образов при наличии в обучающей выборке пропусков данных. На основе анализа условий их асимптотической сходимости определены требования к методам заполнения пропусков данных.

Введение. Одной из проблем теории обучающихся систем является обработка данных с пропусками. Перед исследователем возникает дилемма: отсев информации с пропусками либо их заполнение в соответствии с имеющимися объективными предпосылками.

В первом случае теряются полезная информация, содержащаяся в остальных позициях строки массива обучающей выборки, и материальные затраты на ее получение. Основным направлением в решении проблемы пропуска данных является их восстановление в соответствии с моделями закономерностей взаимосвязи между признаками исходной выборки данных. Особую известность получили локальные алгоритмы заполнения пропусков исходной некомплектной таблицы «объект-признак», основанные на оценивании закономерностей взаимосвязи между ее строками и столбцами в ограниченной окрестности анализируемого элемента [1]. В работе [2] преобразование исходной информации заменяется процедурой «размножения» строк с пропусками данных на основе принципов имитации систем, что позволяет обойти проблему искажений априорных сведений из-за неточности используемых моделей и субъективных предположений исследователя.

Возникающая при этом естественная неоднородность получаемых данных (наличие в преобразованной таблице строк без пропусков и с их заполнением) требует разработки модифицированных непараметрических алгоритмов обработки информации и исследования их свойств.

В данной работе предлагается методика анализа непараметрических алгоритмов распознавания образов в условиях неоднородных выборок, осно-

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант № 07-01-00006) и Совета по грантам Президента РФ (грант № НШ-3431.2008.9).

ванная на исследовании асимптотических свойств оценки уравнения разделяющей поверхности с учетом погрешности используемого метода заполнения пропусков данных.

Непараметрическая оценка плотности вероятности в условиях неоднородных данных. Синтез непараметрических алгоритмов распознавания образов основан на оценивании линейных функционалов от статистических оценок плотности вероятности распределения признаков $x \in R^k$ классифицируемых объектов типа [3, 4]

$$\bar{p}(x) = \left(n \prod_{\nu=1}^{k} c_\nu \right)^{-1} \sum_{i=1}^{n} \prod_{\nu=1}^{k} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu} \right), \qquad (1)$$

восстанавливаемой по выборке $V = (x^i,\ i = \overline{1,n})$. Здесь $\Phi(\cdot)$ – ядерные функции, удовлетворяющие требованиям положительности, симметричности и нормированности, а $c_\nu = c_\nu(n),\ \nu \in J = (\nu = \overline{1,k})$, – последовательности коэффициентов их размытости.

Пусть имеются пропуски данных, и после их заполнения тем или иным методом получаем выборку

$$V1 = (x_\nu^i,\ \nu \in J,\ i \in I \setminus \bar{I};\ x_\nu^i,\ \nu \in J \setminus J^i;\ \bar{x}_\nu^i,\ \nu \in J^i,\ i \in \bar{I}),$$

где $\bar{I}$ – множество номеров ее элементов с заполненными пропусками данных, а $I = (i = \overline{1,n})$.

В выборке $V1$ наблюдения $\bar{x}_\nu,\ \nu \in J^i$ сформированы в соответствии с принятым методом заполнения пропусков данных с погрешностью ε. Полагаем

$$\bar{x}_\nu^i = x_\nu^i + \varepsilon^i, \quad \nu \in J^i,\ i \in \bar{I},$$

где ε^i – наблюдения случайной величины с плотностью вероятности $p(\varepsilon)$. Будем считать $M\varepsilon = 0,\ M\varepsilon^2 = \sigma^2$ (M – знак математического ожидания).

Восстановим плотность вероятности $p(x)$ в виде смеси

$$p(x) = P_1\, p_1(x) + P_2\, p_2(x), \qquad (2)$$

где $P_j,\ j = 1,2,$ – априорные вероятности появления в выборке $V1$ данных без пропусков и с ними.

Для оценивания плотностей вероятности $p_1(x)$ по выборке $V11 = (x_\nu^i,\ \nu \in J,\ i \in I \setminus \bar{I})$ и $p_2(x)$ по данным $V12 = (x_\nu^i,\ \nu \in J \setminus J^i;\ \bar{x}_\nu^i,\ \nu \in J^i,\ i \in \bar{I})$ с заполненными пропусками будем использовать непараметрические статистики ядерного типа (1).

Обозначим через $\bar{n} = |\bar{I}|,\ n_1 = |I \setminus \bar{I}|$ количество элементов множеств $\bar{I}$ и $I \setminus \bar{I}$, причем $n = n_1 + \bar{n}$.

Тогда непараметрическая оценка (2) запишется в виде

$$\bar{\bar{p}}(x) = n^{-1} \left[\sum_{i \in I \setminus \bar{I}} \prod_{\nu \in J} c_\nu^{-1} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu} \right) + \right.$$

$$+\sum_{i\in\bar{I}}\prod_{\nu\in J\setminus J^i} c_\nu^{-1}\Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right)\prod_{\nu\in J^i} c_\nu^{-1}\Phi\left(\frac{x_\nu - \bar{x}_\nu^i}{c_\nu}\right)\Bigg]. \qquad (3)$$

Асимптотические свойства $\bar{\bar{p}}(x)$. Для упрощения выкладок без существенной потери общности получаемых выводов предположим наличие выборки

$$V1=(x_1^i, x_2^i,\ i=\overline{1,n_1};\ \bar{x}_1^i, x_2^i,\ i=\overline{n_1+1,n}),$$

по которой восстанавливается плотность вероятности $p(x_1, x_2)$ с помощью статистики

$$\bar{\bar{p}}(x_1,x_2)=(nc_1c_2)^{-1}\left[\sum_{i=1}^{n_1}\prod_{\nu=1}^{2}\Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right)+\sum_{i=n_1+1}^{n}\Phi\left(\frac{x_1-\bar{x}_1^i}{c_1}\right)\Phi\left(\frac{x_2-x_2^i}{c_2}\right)\right].$$

Предположим, что плотность $p(x_1, x_2)$ ограничена и непрерывна со всеми производными до порядка γ включительно. Эти условия, накладываемые на $p(x_1, x_2)$, обозначим через G_γ.

Справедлива следующая

Теорема. Пусть $p(x)\ \forall\ x\in R^2$ удовлетворяет условиям G_2; ядерные функции $\Phi(u)$ являются положительными, симметричными, нормированными и имеют ограниченные центральные моменты; последовательности коэффициентов размытости $c_1=c_1(n)\geq 0,\ c_2=c_2(n)\geq 0$ ядерных функций таковы, что при $n\to\infty$ значения $c_1\to 0,\ c_2\to 0,\ \bar{n}\sigma^2/n\to 0,\ nc_1c_2\to\infty$. Тогда непараметрическая оценка плотности вероятности $\bar{\bar{p}}(x)$ обладает свойством асимптотической несмещенности и состоятельности.

Доказательство.

1. *Асимптотическая несмещенность*. Определим асимптотическое выражение для математического ожидания $\bar{\bar{p}}(x)$:

$$M\bar{\bar{p}}(x)=\frac{1}{nc_1c_2}\sum_{i=1}^{n_1}M\left[\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{x_2-x_2^i}{c_2}\right)\right]+$$

$$+\frac{1}{nc_1c_2}\sum_{i=n_1+1}^{n}M\left[\Phi\left(\frac{x_1-\bar{x}_1^i}{c_1}\right)\Phi\left(\frac{x_2-x_2^i}{c_2}\right)\right]=$$

$$=\frac{1}{nc_1c_2}\sum_{i=1}^{n_1}\iint\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{x_2-x_2^i}{c_2}\right)p(x_1^i,x_2^i)\,dx_1^i\,dx_2^i+$$

$$+\frac{1}{nc_1c_2}\sum_{i=n_1+1}^{n}\iiint\Phi\left(\frac{x_1-x_1^i-\varepsilon^i}{c_1}\right)\Phi\left(\frac{x_2-x_2^i}{c_2}\right)p(x_1^i,x_2^i)p(\varepsilon^i)\,dx_1^i\,dx_2^i\,d\varepsilon^i. \qquad (4)$$

Бесконечные пределы интегрирования здесь и в дальнейшем опускаются.

Первая сумма после замены переменных преобразуется к виду

$$\frac{n_1}{n}\int\int \Phi(u_1)\Phi(u_2)p(x_1 - c_1u_1,\ x_2 - c_2u_2)\,du_1\,du_2.$$

Разложив в ряд Тейлора плотность вероятности $p(x_1 - c_1u_1, x_2 - c_2u_2)$ в точке $x = (x_1, x_2)$ и проинтегрировав полученное выражение с учетом свойств ядерной функции, имеем

$$\frac{n_1}{n}\left[p(x_1, x_2) + \frac{c_1^2}{2}p^{(2)}_{x_1}(x_1, x_2) + \frac{c_2^2}{2}p^{(2)}_{x_2}(x_1, x_2) + O(c_1^4, c_2^4)\right],$$

где $p^{(2)}_{x_\nu}(x_1, x_2)$ – вторая производная плотности вероятности $p(x_1, x_2)$ по переменной x_ν, $\nu = 1, 2$.

Преобразовав по аналогии вторую сумму выражения (4), получим

$$\frac{\overline{n}}{nc_1c_2}\int\int\int \Phi\left(\frac{x_1 - t_1 - \varepsilon}{c_1}\right)\Phi\left(\frac{x_2 - t_2}{c_2}\right)p(t_1, t_2)p(\varepsilon)\,dt_1\,dt_2\,d\varepsilon \sim$$

$$\sim \frac{\overline{n}}{n}\left[p(x_1, x_2) + \frac{c_1^2}{2}p^{(2)}_{x_1}(x_1, x_2) + \frac{c_2^2}{2}p^{(2)}_{x_2}(x_1, x_2) + \frac{p^{(2)}_{x_1}(x_1, x_2)}{2}\sigma^2 + O(c_1^4, c_2^4)\right].$$

Окончательно имеем

$$M\overline{\overline{p}}(x) \sim p(x_1, x_2) + \frac{c_1^2}{2}p^{(2)}_{x_1}(x_1, x_2) + \frac{c_2^2}{2}p^{(2)}_{x_2}(x_1, x_2) +$$

$$+ \frac{\overline{n}\sigma^2}{n}p^{(2)}_{x_1}(x_1, x_2) + O(c_1^4, c_2^4). \qquad (5)$$

Отсюда при выполнении условий теоремы $c_1(n) \to 0$, $c_2(n) \to 0$ и $\overline{n}\sigma^2 / n \to 0$ с ростом $n \to \infty$ следует асимптотическая несмещенность $\overline{\overline{p}}(x)$. В противном случае будет наблюдаться смещение, которое растет с увеличением дисперсии σ^2 погрешности ε метода заполнения пропусков данных и их количества $\overline{n}$.

2. *Сходимость в среднеквадратическом*. Рассмотрим выражение

$$M(\overline{\overline{p}}(x) - p(x))^2 = M\overline{\overline{p}}^2(x) - 2p(x)M\overline{\overline{p}}(x) + p^2(x), \qquad (6)$$

где дополнительного исследования требует составляющая

$$M\overline{\overline{p}}^2(x) = \frac{1}{n^2c_1^2c_2^2}\left[\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} M\left(\Phi\left(\frac{x_1 - x_1^i}{c_1}\right)\Phi\left(\frac{x_2 - x_2^i}{c_2}\right)\Phi\left(\frac{x_1 - x_1^j}{c_1}\right)\Phi\left(\frac{x_2 - x_2^j}{c_2}\right)\right) +\right.$$

$$+\sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n} M\left(\Phi\left(\frac{x_1-\bar{x}_1^{\,i}}{c_1}\right)\Phi\left(\frac{x_2-x_2^{i}}{c_2}\right)\Phi\left(\frac{x_1-\bar{x}_1^{\,j}}{c_1}\right)\Phi\left(\frac{x_2-x_2^{j}}{c_2}\right)\right)+$$

$$+2\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} M\left(\Phi\left(\frac{x_1-x_1^{i}}{c_1}\right)\Phi\left(\frac{x_2-x_2^{i}}{c_2}\right)\Phi\left(\frac{x_1-\bar{x}_1^{\,j}}{c_1}\right)\Phi\left(\frac{x_2-x_2^{j}}{c_2}\right)\right)\Bigg].$$

Последовательно анализируя слагаемые $M\bar{\bar{p}}^{\,2}(x)$ в соответствии с вышеприведенной технологией при достаточно больших n, нетрудно получить

$$M\bar{\bar{p}}^{\,2}(x) = p^2(x_1,x_2) + \frac{(\int\Phi^2(u)\,du)^2\,p(x_1,x_2)}{nc_1c_2} +$$

$$+\frac{1}{4}(c_1^2 p_{x_1}^{(2)}(x_1,x_2)+c_2^2 p_{x_2}^{(2)}(x_1,x_2))^2 +$$

$$+\frac{\sigma^2\bar{n}\,p_{x_1}^{(2)}(x_1,x_2)}{2n}\left[c_1^2 p_{x_1}^{(2)}(x_1,x_2)+\frac{n_1+\bar{n}c_1^2}{n}+c_2^2 p_{x_2}^{(2)}(x_1,x_2)+\frac{\sigma^2\bar{n}}{2n}\,p_{x_1}^{(2)}(x_1,x_2)\right]+$$

$$+c_1^2\,p(x_1,x_2)p_{x_1}^{(2)}(x_1,x_2)+c_2^2\,p(x_1,x_2)p_{x_2}^{(2)}(x_1,x_2)+$$

$$+\frac{\bar{n}\sigma^2}{n}\,p_{x_1}^{(2)}(x_1,x_2)p(x_1,x_2)+O(c_1^4,c_2^4)+O(c_1^2,c_2^2)+O\left(\frac{1}{n}\right). \qquad (7)$$

Подставим полученный результат (7) с учетом $M\bar{\bar{p}}(x)$ (4) в выражение (6) и после несложных преобразований будем иметь

$$M(\bar{\bar{p}}(x)-p(x))^2 \sim \frac{(\int\Phi^2(u)\,du)^2\,p(x_1,x_2)}{nc_1c_2} +$$

$$+\frac{1}{4}(c_1^2 p_{x_1}^{(2)}(x_1,x_2)+c_2^2 p_{x_2}^{(2)}(x_1,x_2))^2+\frac{\sigma^2\bar{n}\,p_{x_1}^{(2)}(x_1,x_2)}{2n}\times$$

$$\times\left[c_1^2 p_{x_1}^{(2)}(x_1,x_2)+\frac{n_1+\bar{n}c_1^2}{n}+c_2^2 p_{x_2}^{(2)}(x_1,x_2)+\frac{\sigma^2\bar{n}}{2n}\,p_{x_1}^{(2)}(x_1,x_2)-2p(x_1,x_2)\right]. \qquad (8)$$

Отсюда следует, что $\bar{\bar{p}}(x)$ обладает свойством сходимости в среднеквадратическом, а с учетом свойств асимптотической несмещенности (5) является состоятельной, если выполняются условия $nc_1c_2\to\infty$, $c_1\to 0$, $c_2\to 0$, $\bar{n}\sigma^2/n\to 0$ при $n\to\infty$.

Непараметрические алгоритмы распознавания образов в условиях пропуска данных. Пусть $V = \bigcup_{j=1}^{M} V_j$ – неоднородная обучающая выборка при решении многоальтернативной задачи распознавания образов. Элементы выборки *j*-го класса

$$V_j = (x_\nu^i,\ \nu \in J,\ \ i \in I_j \setminus \bar{I}_j;\ \ x_\nu^i,\ \nu \in J \setminus J^i;\ \ \bar{x}_\nu^i,\ \nu \in J^i;\ \ i \in \bar{I}_j)$$

составлены из признаков x_ν^i, $\nu \in J = (\nu = \overline{1,k})$, классифицируемых объектов без пропусков и с заполненными пропусками данных $\bar{x}_\nu^i$, $\nu \in J^i \subset J$, $i \in \bar{I}_j \subset \subset I_j$. Здесь I_j – множество номеров элементов обучающей выборки V, принадлежащих *j*-му классу. Обозначим через n_j и $\bar{n}_j$ количество элементов множеств $I_j, \bar{I}_j$ соответственно.

Воспользуемся статистикой типа (3) при построении непараметрической оценки плотности вероятности распределения признаков x в *j*-м классе:

$$\bar{\bar{p}}_j(x) = n_j^{-1}\left[\sum_{i \in I_j \setminus \bar{I}_j} \prod_{\nu \in J} c_\nu^{-1} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right) + \right.$$

$$\left. + \sum_{i \in \bar{I}_j} \prod_{\nu \in J \setminus J^i} c_\nu^{-1} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right) \prod_{\nu \in J^i} c_\nu^{-1} \Phi\left(\frac{x_\nu - \bar{x}_\nu^i}{c_\nu}\right)\right]. \qquad (9)$$

Тогда оценка решающего правила распознавания образов, соответствующего, например, критерию максимального правдоподобия, запишется в виде

$$\bar{m}(x):\ x \in \Omega_j,\ \ \text{если}\ \ \bar{\bar{p}}_j(x) = \max_{t = \overline{1,\ M}} \bar{\bar{p}}_t(x). \qquad (10)$$

Оптимизация $\bar{m}(x)$ по коэффициентам размытости c_ν, $\nu = \overline{1,k}$, ядерных функций осуществляется из условия минимума эмпирической ошибки классификации в режиме «скользящего экзамена»:

$$\bar{\rho}(x) = n_\nu^{-1} \sum_{j=1}^{M} \sum_{i \in I_j} 1(\sigma(i),\ \bar{\sigma}(i)),$$

где n_ν – объем обучающей выборки V; $\sigma(i)$ – «указание учителя» о принадлежности ситуации x^i к одному из M классов; $\bar{\sigma}(i)$ – «решение» правила (10);

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

Асимптотические свойства непараметрической решающей функции. Для использования результатов предыдущего раздела сохраним условия и предположения, введенные при исследовании асимптотических свойств непараметрической оценки плотности вероятности $\bar{\bar{p}}(x)$.

Рассмотрим двухальтернативную задачу распознавания образов. В этом случае непараметрическая оценка уравнения разделяющей поверхности, соответствующая решающему правилу (10), представима в виде

$$\bar{f}_{12}(x) = \bar{\bar{p}}_1(x) - \bar{\bar{p}}_2(x). \tag{11}$$

Ее составляющие $\bar{\bar{p}}_1(x)$ и $\bar{\bar{p}}_2(x)\ \forall\ x \in R^2$ восстанавливаются на основе статистики типа (3) при размерности вектора x, равной 2, по выборкам:

$$V_1 = (x_1^i,\ x_2^i,\ i = \overline{1, n_1^1};\ \ \bar{x}_1^i,\ x_2^i,\ i = \overline{n_1^1 + 1, n_1}),$$

$$V_2 = (x_1^i,\ x_2^i,\ i = \overline{1, n_2^1};\ \ \bar{x}_1^i,\ x_2^i,\ i = \overline{n_2^1 + 1, n_2}).$$

Так как $\bar{f}_{12}(x)$ является линейным функционалом от $\bar{\bar{p}}_1(x)$, $\bar{\bar{p}}_2(x)$, то в соответствии с (5) непараметрическая оценка уравнения разделяющей поверхности (11) обладает свойством асимптотической несмещенности.

Примем в качестве основного показателя эффективности $\bar{f}_{12}(x)$ среднеквадратический критерий

$$W = M(\bar{f}_{12}(x) - f_{12}(x))^2, \tag{12}$$

отражающий точность аппроксимации $\bar{f}_{12}(x)$ байесовской решающей функции

$$f_{12}(x) = p_1(x) - p_2(x). \tag{13}$$

С учетом (11) и (13) представим (12) в виде

$$W(x) = M(p_1(x) - \bar{\bar{p}}_1(x))^2 + M(p_2(x) - \bar{\bar{p}}_2(x))^2 -$$

$$-2M[(p_1(x) - \bar{\bar{p}}_1(x))(p_2(x) - \bar{\bar{p}}_2(x))]. \tag{14}$$

На основе результатов (5) и (8) определим асимптотическое выражение (12). Пренебрегая величинами малости $(c_\nu^2 \bar{n}_j)/n_j$, $\nu = 1, 2$, $j = 1, 2$, где $\bar{n}_j = n_j - n_j^1$, получим

$$\overline{W}(x) \sim \sum_{j=1}^{2} \left[\frac{(\int \Phi^2(u)\,du)^2}{n_j c_1 c_2}\ p_j(x) + \frac{\sigma^2 \bar{n}_j}{2n_j}\ p_{j,1}^{(2)}(x) \times \right.$$

$$\left. \times \left(n_j^{-1} n_j^1 + \frac{\sigma^2 \bar{n}_j}{2n_j}\ p_{j,1}^{(2)}(x) - 2p_j(x) \right) \right] - \frac{2\bar{n}_1 \bar{n}_2}{n_1 n_2}\ \sigma^4 p_{1,1}^{(2)}(x) p_{2,1}^{(2)}(x) +$$

$$+ \frac{1}{4}[c_1^2(p_{1,1}^{(2)}(x) - p_{2,1}^{(2)}(x)) + c_2^2(p_{1,2}^{(2)}(x) - p_{2,2}^{(2)}(x))]^2,$$

где $p^{(2)}_{j,1}(x)$, $p^{(2)}_{j,2}(x)$ – вторые производные по x_1, x_2 плотностей вероятности $p_j(x)$, $j=1,2$; $\overline{n}_j$ – количество заполненных пропусков измерений x_1 в j-м классе.

Для удобства последующего анализа получаемых результатов будем считать $n_j = n$, $\overline{n}_j = \overline{n}$, $j=1, 2,$ и $c_1 = c_2 = c$. Тогда интеграл от $\overline{W}(x)$ перепишется в виде

$$\overline{\overline{W}} \sim \frac{2(\int \Phi^2(u)\,du)^2}{nc^2} + \frac{c^4}{4}\beta + \frac{\sigma^2\overline{n}}{n}\left(\frac{n-\overline{n}}{2n}\beta_1 - \beta_3\right) + \frac{\sigma^4\overline{n}^2}{n^2}\left(\frac{\beta_2}{4} - 2\beta_4\right), \qquad (15)$$

где

$$\beta = \int\int ((p^{(2)}_{1,1}(x) - p^{(2)}_{2,1}(x)) + (p^{(2)}_{1,2}(x) - p^{(2)}_{2,2}(x)))^2 dx_1 dx_2;$$

$$\beta_1 = \int\int (p^{(2)}_{1,1}(x) + p^{(2)}_{2,1}(x))\,dx_1 dx_2;$$

$$\beta_2 = \int\int (p^{(2)}_{1,1}(x) - p^{(2)}_{2,1}(x))^2 dx_1 dx_2;$$

$$\beta_3 = \int\int (p^{(2)}_{1,1}(x)p_1(x) + p^{(2)}_{2,1}(x)p_2(x))\,dx_1 dx_2;$$

$$\beta_4 = \int\int p^{(2)}_{1,1}(x)p^{(2)}_{2,1}(x)\,dx_1 dx_2.$$

Нетрудно заметить, что непараметрическая оценка $\bar{f}_{12}(x)$ (11) уравнения разделяющей поверхности $f_{12}(x)$ (13) обладает свойством сходимости в среднеквадратическом, если при $n \to \infty$ выполняются условия $nc^2 \to \infty$, $c \to 0$, $\overline{n}/n \to 0$. Поэтому с учетом асимптотической несмещенности $\bar{f}_{12}(x)$ является состоятельной оценкой $f_{12}(x)$.

Анализ аппроксимационных свойств $\bar{f}_{12}(x)$. Определив оптимальный коэффициент размытости ядерной функции для $\bar{f}_{12}(x)$ из условия минимума по c выражения $\overline{\overline{W}}$ (15), получим

$$c^* = \left[\frac{2(\int \Phi^2(u)\,du)^2}{n\beta}\right]^{1/6}.$$

При $c = c^*$ минимальное значение асимптотического выражения $\overline{\overline{W}}$ среднеквадратического критерия равно

$$\overline{\overline{W}}^* = \frac{3}{2}\left(\frac{(\int \Phi^2(u)\,du)^4\beta}{2n^2}\right)^{1/3} + \frac{\sigma^2\overline{n}}{n}\left(\frac{n-\overline{n}}{2n}\beta_1 - \beta_3\right) + \frac{\sigma^4\overline{n}^2}{n^2}\left(\frac{\beta_2}{4} - 2\beta_4\right). \qquad (16)$$

Оценим значимость информации, содержащейся в элементах обучающей выборки с заполненными пропусками данных. Для этого рассмотрим от-

ношение $D = \overline{\overline{W}}^* \big/ \widetilde{W}^*$ значения $\overline{\overline{W}}^*$ (16) к соответствующему значению $\widetilde{W}^*$ для непараметрической решающей функции, восстанавливаемой по обучающей выборке $\widetilde{V} = (x^i, \sigma(i),\ i = \overline{1,2n})$ без пропусков данных. В принятых условиях $\widetilde{W}^*$ соответствует первому слагаемому в выражении (16) [5].

Тогда при соблюдении принятого ранее условия $\overline{n}/n \to 0 \,\forall\, n \to \infty$ отношение

$$D \sim 1 + \frac{\sigma^2 \overline{n}}{2\left(\frac{1}{2}\left(\int \Phi^2(u)\,du\right)^4 \beta n\right)^{1/3}} \left(\frac{\beta_1}{2} - \beta_3\right) \qquad (17)$$

имеет предел, равный единице, если $\overline{n} \big/ n^{1/3} \to 0$ при $n \to \infty$.

Тем самым обосновывается существование условий, когда заполнение пропусков данных позволяет восполнить информацию обучающей выборки, достаточной для эффективного решения задачи классификации с помощью непараметрических методов.

Нетрудно показать, что даже при постоянной дисперсии σ^2 погрешности метода заполнения пропусков данных приведенные выше требования выполняются при количестве ситуаций в обучающей выборке с пропусками данных $\overline{n} < \lambda \sqrt[3]{n}$.

Применение методов заполнения пропусков данных, обеспечивающих уменьшение дисперсии $\sigma^2(n) \to 0$ с ростом объема выборки $n \to \infty$, позволяет значительно повысить эффективность непараметрических алгоритмов классификации. Положив $\overline{n} = \lambda n^{\alpha}$, $\alpha < 1$, $\sigma^2(n) = \lambda_1 n^{-\gamma}$, можно определить, что порядок асимптотической сходимости второго слагаемого к нулю достигает уровня $n^{-(1/3 + \gamma - \alpha)}$. Отсюда следует очевидное ограничение на $\alpha < \frac{1}{3} + \gamma$ и на возможное количество пропусков $\overline{n} < \lambda_1 n^{1/3 + \gamma}$ в исходной обучающей выборке.

Заключение. Непараметрические алгоритмы распознавания образов, синтез которых осуществляется на основе обучающих выборок с пропусками данных, обладают свойством асимптотической сходимости. При конечной дисперсии погрешности метода заполнения пропусков данных существуют условия асимптотической несмещенности и состоятельности непараметрической оценки уравнения разделяющей поверхности, основанной на ядерной оценке плотности вероятности типа Розенблатта – Парзена. Снижение дисперсии с ростом объема обучающей выборки способствует повышению эффективности непараметрических алгоритмов классификации. Обосновано существование предела, равного единице, для отношения среднеквадратических критериев точности аппроксимации байесовского уравнения разделяющей поверхности ее непараметрическими оценками, построенными по выборкам с заполненными пропусками данных и без них. С этих позиций определены условия, когда заполнение пропусков данных позволяет восполнить информацию обучающей выборки, достаточной для эффективного решения задачи классификации с помощью непараметрических методов.

СПИСОК ЛИТЕРАТУРЫ

1. **Загоруйко Н. Г., Елкина В. Н., Тимеркаев В. С.** Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм ZET) // Вычислительные системы «Эмпирическое предсказание и распознавание образов». Новосибирск: Изд-во ИМ СО АН СССР, 1975. Вып. 61. С. 3.
2. **Лапко А. В., Лапко В. А., Цугленок Г. И.** Синтез и анализ непараметрических моделей стохастических зависимостей и распознавания образов в условиях пропуска данных // Вест. КрасГАУ. 2005. № 7. С. 64.
3. **Parzen E.** On the estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**. P. 1065.
4. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. **14**, вып. 1. С. 156.
5. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000.

Поступила в редакцию 13 февраля 2007 г.