РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2007, том 43, № 5

УДК 681.513

НЕПАРАМЕТРИЧЕСКИЕ АЛГОРИТМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ ПРИ СЛУЧАЙНЫХ ЗНАЧЕНИЯХ КОЭФФИЦИЕНТОВ РАЗМЫТОСТИ ЯДЕРНЫХ ФУНКЦИЙ*

А. В. Лапко, В. А. Лапко

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

Предлагаются непараметрические алгоритмы распознавания образов, основанные на рандомизированном методе их оптимизации. Идея рассматриваемого подхода состоит в признании случайного характера коэффициентов размытости ядерных функций и выборе параметров закона их распределения при оптимизации непараметрических решающих правил. Исследуются свойства разработанных классификаторов и анализируются результаты их сравнения с традиционными непараметрическими алгоритмами распознавания образов.

Введение. Парадокс традиционных методов идентификации стохастических моделей состоит в сопоставлении конечной случайной выборки наблюдений переменных изучаемых объектов с конкретным набором параметров модели, оптимальных в некотором смысле. Предлагается принципиально новый рандомизированный подход определения коэффициентов размытости непараметрических алгоритмов распознавания образов, основанных на ядерной оценке плотности вероятности типа Розенблатта – Парзена.

Впервые методика случайного выбора коэффициентов размытости ядерных функций при синтезе непараметрической оценки плотности вероятности была предложена в 1975 году Т. Вагнером [1]. Формирование случайной последовательности коэффициентов размытости при оценивании плотности вероятности $p(x)$ осуществляется из выборки расстояний между исходными наблюдениями $(x^i,\ i=\overline{1,n})$ и их k-ми ближайшими соседями. Несмотря на кажущуюся простоту подхода, остается открытой проблема выбора значения k и обоснование последствий такого выбора.

Цель данной работы – на основе анализа асимптотических свойств непараметрической оценки $\overline{p}(x)$ плотности вероятности $p(x)$ показать возможности нахождения рационального закона распределения $p(c)\ \forall\ c\in(0;\ h]$ коэффициентов размытости c в классе степенных функций и использования полученных результатов при синтезе непараметрических алгоритмов распознавания образов в условиях случайных значений коэффициентов размыто-

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант № 07-01-00006) и ведущих научных школ (грант № НШ3428.2006.9).

сти ядерных функций, а также возможность исследования их свойств методом статистического моделирования.

Рандомизированный метод оптимизации. Обоснование рандомизированного метода оптимизации непараметрических алгоритмов распознавания образов рассмотрим на примере оценивания плотности вероятности. Пусть $V = (x^i,\ i = \overline{1,n})$ – выборка из n статистически независимых наблюдений случайной величины $x \in R^1$ с плотностью вероятности $p(x)$, вид которой неизвестен. Будем считать, что $p(x)$ ограничена и непрерывна со всеми своими производными до второго порядка включительно. В качестве приближения по эмпирическим данным V искомой плотности вероятности $p(x)$ примем статистику типа Розенблатта – Парзена [2]

$$\overline{p}(x) = (nc)^{-1} \sum_{i=1}^{n} \Phi\left(\frac{x - x^i}{c}\right), \tag{1}$$

где $\Phi(\cdot)$ – ядерные функции, удовлетворяющие условиям положительности, симметричности и нормированности; $c = c(n)$ – последовательность положительных чисел (коэффициентов размытости) таких, что

$$\lim_{n\to\infty} c(n) = 0, \qquad \lim_{n\to\infty} nc(n) = \infty.$$

Свойства непараметрической оценки плотности вероятности (1): асимптотическая несмещенность, состоятельность, сходимость почти наверное к $p(x)$ – подробно исследованы в работах [3, 4]. Показано, что среднеквадратическая ошибка аппроксимации

$$\overline{W}(c) = M\left(\int [p(x) - \overline{p}(x)]^2\, dx\right) = M\left\| p(x) - \overline{p}(x)\right\|^2 \sim$$

$$\sim (nc)^{-1} \int \Phi^2(x)\, dx + \frac{c^4 \left\| p^{(2)}(x)\right\|^2}{4}. \tag{2}$$

Здесь $p^{(2)}(x)$ – вторая производная $p(x)$ по x.

Исходя из случайного характера коэффициентов размытости ядерных функций в (1), будем искать рациональный закон их распределения среди функций вида

$$p_h(c) = \alpha c^t, \qquad \alpha = \frac{t+1}{h^{t+1}} \ \forall\ c \in (0;\ h], \tag{3}$$

которые не противоречат условиям асимптотической сходимости непараметрических статистик. Параметр t плотности вероятности $p_h(c)$ априори не определен.

Левая граница интервала изменения c следует из условий асимптотической несмещенности $\overline{p}(x)$: $c(n) \to 0$ при увеличении объема исходных данных $n \to \infty$.

Вычислим асимптотическое выражение среднеквадратического отклонения $\overline{p}(x)$ от $p(x)$ при случайных значениях коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности $\overline{p}(x)$:

$$\overline{W}_p(h) = \int_0^h \overline{W}(c)\, p(c)\, dc \sim \frac{t+1}{t}(nh)^{-1} \int \Phi^2(x)\, dx + \frac{t+1}{5+t} \frac{h^4 \left\| p^{(2)}(x) \right\|^2}{4}. \quad (4)$$

Для сравнения традиционного и рандомизированного методов оптимизации непараметрической оценки плотности вероятности определим отношение $\overline{W}(c^*) \big/ \overline{W}_p(h^*)$ соответствующих им асимптотических выражений среднеквадратических критериев при оптимальных параметрах $c = c^*$ и $h = h^*$.

Из условия минимума $\overline{W}(c)$ и $\overline{W}_p(h)$ по c и h нетрудно получить

$$c^* = \left[\frac{\int \Phi^2(x)\, dx}{n \left\| p^{(2)}(x) \right\|^2} \right]^{1/5}, \qquad h^* = \left(\frac{5+t}{t} \right)^{1/5} c^*.$$

При оптимальных параметрах c^*, h^* отношение

$$\frac{\overline{W}(c^*)}{\overline{W}_p(h^*)} \sim \frac{(1+5t^{-1})^{1/5}}{1+t^{-1}} \quad (5)$$

меньше единицы при конечных значениях параметра t плотности вероятности $p_h(c)$ коэффициентов размытости ядерных функций.

Однако использование непараметрической оценки со случайными значениями коэффициентов размытости ядерных функций

$$\widetilde{p}(x) = n^{-1} \sum_{i=1}^{n} \frac{1}{c^i} \Phi\left(\frac{x - x^i}{c^i} \right) \quad (6)$$

позволяет снизить ее смещение при оценивании плотности вероятности по сравнению с традиционной статистикой типа (1).

Можно показать, что асимптотическое выражение смещения $\widetilde{p}(x)$ от $p(x)$ имеет вид

$$\overline{W}_p^1(h) = \int_0^h M(\widetilde{p}(x) - p(x))\, p(c)\, dc \sim \frac{h^2(t+1)}{2(t+3)} p^{(2)}(x), \quad (7)$$

а его отношение к смещению $\overline{W}^1(c)$ для традиционной непараметрической оценки плотности вероятности $\overline{p}(x)$ при оптимальных параметрах c^* и h^* равно

$$\frac{\overline{W}_p^1(h^*)}{\overline{W}^1(c^*)} \sim \frac{t+1}{t+3} \left(\frac{5+t}{t} \right)^{2/5}. \quad (8)$$

Если параметр t плотности вероятности $p_h(c)$ больше или равен двум, то отношение (8) меньше единицы.

Анализ выражений (4), (7) показывает, что непараметрическая оценка плотности вероятности $\widetilde{p}(x)$ со случайными значениями коэффициентов размытости (6) обладает свойствами асимптотической несмещенности и состоятельности. Она характеризуется более низким значением смещения (7) и несколько большим значением среднеквадратического отклонения (4) по сравнению с непараметрической статистикой (1). Следует ожидать проявления потенциальной эффективности непараметрической оценки плотности вероятности (6) при конечных объемах статистических данных.

Модифицированный непараметрический алгоритм классификации. Определим коэффициенты размытости ядерных функций в виде $c_{\nu} = c\overline{\sigma}_{\nu}$, где $\overline{\sigma}_{\nu}$ – оценка среднеквадратических отклонений параметров x_{ν}, $\nu = \overline{1, k}$, классифицируемых объектов, а c – случайная величина с плотностью вероятности $p_h(c)\ \forall\ c \in (0;\ h]$.

Примем процедуру формирования последовательности параметров

$$c = h\varepsilon^{1/(t+1)} \tag{9}$$

на основании случайной величины $\varepsilon \in [0; 1]$ с равномерным законом распределения. Она может быть получена в результате решения уравнения [5]

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

Сформируем на основании процедуры (9) последовательность коэффициентов размытости и сопоставим случайным образом ее элементам ядерные функции в непараметрических оценках плотностей вероятности байесовского уравнения разделяющей поверхности, соответствующего критерию максимума апостериорной вероятности [6]. Тогда непараметрическая оценка уравнения разделяющей поверхности со случайными коэффициентами размытости ядерных функций для двухальтернативной задачи распознавания образов запишется в виде

$$\widetilde{f}_{12}(x) = \frac{1}{n\prod\limits_{\nu=1}^{k}\overline{\sigma}_{\nu}} \sum_{i=1}^{n} \sigma(i) \prod_{\nu=1}^{k} \frac{1}{c^i} \Phi\left(\frac{x_{\nu} - x_{\nu}^i}{c^i\,\overline{\sigma}_{\nu}}\right),$$

где $\sigma(i)$ – «указания учителя» из обучающей выборки $(x^i,\ \sigma(i),\ i = \overline{1, n})$:

$$\sigma(i) = \begin{cases} -1\ \forall\ x^i \in \Omega_1, \\ \ \ 1\ \forall\ x^i \in \Omega_2. \end{cases}$$

Оптимизация непараметрического алгоритма распознавания образов

$$\widetilde{m}_{12}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \widetilde{f}_{12}(x) \le 0, \\ x \in \Omega_2, & \text{если } \widetilde{f}_{12}(x) > 0, \end{cases} \tag{10}$$

по правой границе h области определения плотности вероятности $p_h(c)$ осуществляется из условия минимума эмпирической ошибки классификации методом «скользящего экзамена»

$$\rho(h) = \frac{1}{n}\sum_{j=1}^{n} 1(\sigma(j), \tilde{\sigma}(j)),$$

где

$$1(\sigma(j), \tilde{\sigma}(j)) = \begin{cases} 1, & \text{если } \sigma(j) \neq \tilde{\sigma}(j), \\ 0, & \text{если } \sigma(j) = \tilde{\sigma}(j); \end{cases}$$

$\tilde{\sigma}(j)$ – «решение» алгоритма (10) о принадлежности объекта, характеризуемого набором признаков x^j, к одному из двух классов.

В многоальтернативной задаче распознавания образов при наличии классов Ω_j, $j = \overline{1,q}$, классификация объектов осуществляется в соответствии с методом дихотомии на основе последовательности решающих правил типа (10) либо используется непараметрический алгоритм

$$\tilde{m}(x)\colon x \in \Omega_j, \text{ если } \overline{P}_j \tilde{p}_j(x) = \max_{t=\overline{1,q}} \overline{P}_t \tilde{p}_t(x), \qquad (11)$$

где $\tilde{p}_t(x)$ – ядерные оценки плотностей вероятности $x \in \Omega_t$ со случайными значениями коэффициентов размытости, а $\overline{P}_t$ – оценки априорных вероятностей появления ситуации x в классах Ω_t, $t = \overline{1,q}$.

Коллектив непараметрических алгоритмов классификации. Используем принципы синтеза коллектива решающих правил для повышения эффективности непараметрических алгоритмов распознавания образов в условиях случайных значений коэффициентов размытости ядерных функций. Пусть $\tilde{m}_{12}^j(x)$, $j = \overline{1, M}$, – непараметрические решающие правила для двухальтернативной задачи распознавания образов, которые построены по одной и той же обучающей выборке $V = (x^i, \sigma(i),\ i = \overline{1,n})$ в соответствии с вышеизложенной методикой. Решающие правила характеризуются одним и тем же оптимальным параметром h правой границы области определения плотности вероятности $p_h(c)$ коэффициента размытости, но разными их случайными последовательностями $(c_j^i,\ i = \overline{1,n})$, $j = \overline{1, M}$.

Воспользуемся одним из известных подходов коллективного оценивания [7], например методом «голосования», и построим решающее правило

$$\tilde{\tilde{m}}_{12}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } M_1/M \geq M_2/M, \\ x \in \Omega_2, & \text{если } M_1/M < M_2/M, \end{cases} \qquad (12)$$

где M_j, $j = 1,2,$ – число «решений», которые принимают члены коллектива, о принадлежности объекта с набором признаков x в пользу j-го класса.

В многоальтернативной постановке задачи распознавания образов каждый член коллектива $\tilde{m}_{12}^j(x)$, $j = \overline{1, M}$, использует решающее правило типа (12). Окончательный вывод, например $x \in \Omega_t$, принимается, если частота решений членов коллектива в пользу t класса максимальная.

Применение коллектива (12) позволяет повысить достоверность принимаемых решений в условиях случайных значений коэффициентов размытости непараметрических алгоритмов.

Результаты вычислительного эксперимента. Исследования осуществлялись при решении двухальтернативной задачи распознавания образов в k-мерном пространстве признаков. Законы распределения признаков классифицируемых объектов в области первого класса формировались в соответствии с датчиками случайных чисел:

$$x_{\nu} = a + \varepsilon(b-a), \quad x_{\nu+1} = (x_{\nu})^2 - 6x_{\nu} + 10 + \sigma_1 \left(\sum_{i=1}^{p_1} \varepsilon^i - 0{,}5\, p_1 \right) \frac{6}{\sqrt{3p_1}}, \ \nu \in I_{\text{н}},$$

где $a = 1{,}5$; $b = 4{,}5$; $p_1 = 5$; среднеквадратическое отклонение $\sigma_1 = 0{,}7$; $\varepsilon \in \in [0; 1]$ – случайная величина с равномерным законом распределения; $I_{\text{н}} = = (1, 3, 5, \ldots)$ – множество нечетных чисел, меньших k.

Признаки второго класса генерировались с нормальным законом распределения

$$x_{\nu} = m + \sigma_{\nu} \left(\sum_{i=1}^{p_2} \varepsilon^i - 0{,}5\, p_2 \right) \frac{6}{\sqrt{3p_2}}, \quad \nu = \overline{1, k},$$

при $p_2 = 5$, $m = \underline{3,} \sigma_{\nu} = 0{,}7$ для нечетных и $\sigma_{\nu} = 0{,}9$ для четных номеров признаков x_{ν}, $\nu = 1, k$.

Рассматриваемые законы распределения признаков в классах для двумерного случая иллюстрирует рис. 1. Здесь и в дальнейшем априорные вероятности классов $P_1 = P_2 = 0{,}5$, т. е. количество наблюдений ситуаций первого (n_1) и второго (n_2) классов в обучающей выборке $(x^i, \sigma(i), \ i = 1, n)$ равны.

Вычислительные эксперименты при фиксированных условиях исследования осуществлялись N раз. Контрольная выборка при оценивании ошибок классификации непараметрических алгоритмов составляла $n_k = 2000$ наблюдений. Достоверность различия оценок ошибок сравниваемых алгорит-

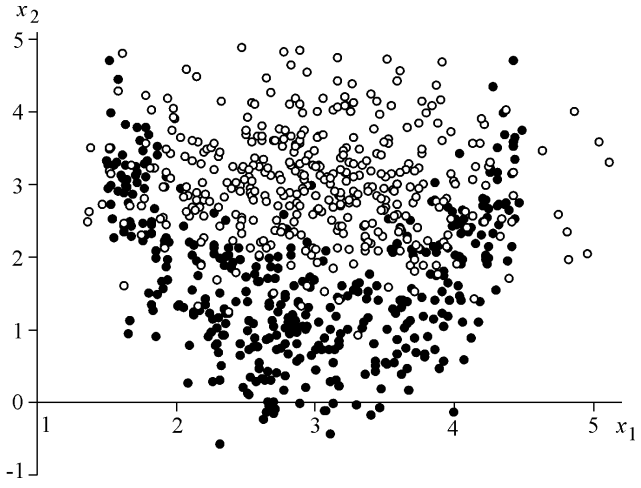

Рис. 1. Элементы обучающей выборки в пространстве двух признаков x_1, x_2 при $n = 600$ (ситуации первого (●) и второго (○) классов)

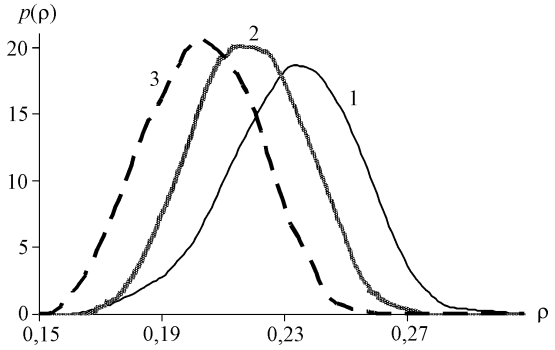

Рис. 2. Оценки плотностей вероятности $p(\rho)$ эмпирических ошибок распознавания образов ρ традиционного непараметрического классификатора (кривая 1), его модификации со случайными значениями коэффициентов размытости (кривая 2) и коллектива (12) при $M = 15$ (кривая 3). Условия эксперимента: объем обучающей выборки $n = 100$; размерность пространства признаков $k = 2$; параметр t закона распределения c (3) равен единице; количество вычислительных экспериментов $N = 100$

мов распознавания образов и законов их распределения рассчитывалась в соответствии с критерием Смирнова.

При синтезе рассматриваемых непараметрических классификаторов использовались параболические ядерные функции из [3].

В данных условиях исследовались свойства традиционного непараметрического классификатора, его модификации со случайными значениями коэффициентов размытости и их коллективов в зависимости от размерности пространства признаков x_v, $v = \overline{1,k}$, и параметров t закона распределения $p_h(c)$ (3) при относительно малых объемах обучающих выборок ($n = 100$).

Коллектив традиционных непараметрических алгоритмов распознавания образов формировался на основе решающего правила (12). Синтез его элементов осуществлялся по обучающим выборкам объема $n_T = \gamma n$, $\gamma < 1$, получаемым случайным образом из исходной выборки $(x^i, \sigma(i), i = \overline{1,n})$. При организации вычислительных экспериментов параметр γ принимался равным 0,9.

Анализ результатов вычислительных экспериментов показывает, что статистические оценки законов распределения эмпирических ошибок распознавания образов традиционного непараметрического алгоритма классификации и его модификации (10) достоверно не отличаются. Однако возможная неточность при определении оценок оптимальных значений коэффициентов размытости приводит к снижению эффективности традиционного непараметрического алгоритма распознавания образов по сравнению с его модификацией.

Законы распределения $p(\rho)$ оценок вероятностей ошибок распознавания образов ρ исследуемых непараметрических алгоритмов классификации при параметрах $c = 1{,}2c^*$, $h = 1{,}2h^*$ коэффициентов размытости ядерных функций, несколько отличающихся от оптимальных c^*, h^*, приведены на рис. 2.

Существенно большая устойчивость эффективности непараметрического алгоритма со случайными коэффициентами размытости от изменения параметра h иллюстрируется рис. 3.

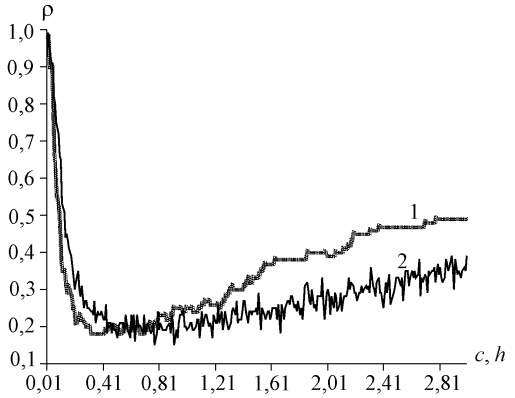

Рис. 3. Зависимости оценок ошибки распознавания ρ традиционного непараметрического алгоритма распознавания образов при $n = 100$ от параметра c коэффициента размытости (кривая 1) и его модификации (10) (кривая 2) от значений параметра h закона распределения (3) при $t = 1$

Потенциальные возможности рандомизированного метода оптимизации наиболее полно реализуются при использовании коллектива (12) непараметрических алгоритмов распознавания образов (10). Наблюдается достоверное преимущество решающего правила (12) при количестве его элементов $M > 5$ перед традиционным непараметрическим классификатором, их коллективом и алгоритмом распознавания образов (10) со случайными коэффициентами размытости ядерных функций.

С увеличением размерности k пространства признаков классифицируемых объектов при фиксированном объеме обучающей выборки n снижение эффективности сравниваемых непараметрических алгоритмов распознавания образов не наблюдается (рис. 4). Однако преимущество коллектива непараметрических решающих правил со случайными коэффициентами размы-

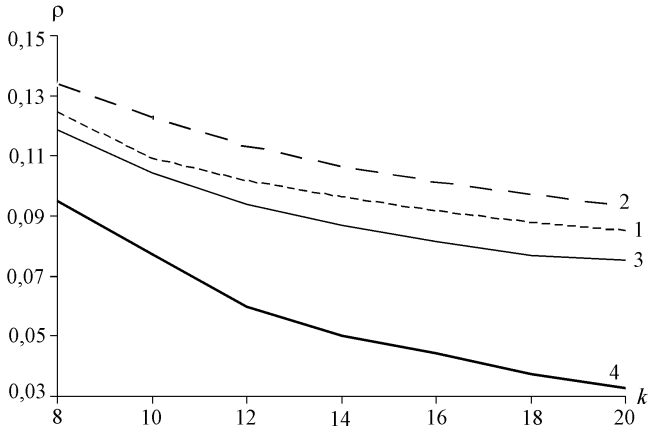

Рис. 4. Зависимости оценок ошибок распознавания образов ρ от размерности k пространства признаков для традиционного непараметрического классификатора (кривая 1) и его коллектива (кривая 3), непараметрического алгоритма (10) со случайными значениями коэффициентов размытости (кривая 2) и его коллектива (кривая 4). Условия эксперимента: $n = 100$; $N = 50$; параметр t закона распределения c (3) равен единице; количество элементов исследуемых коллективов $M = 15$; объем контрольной выборки $n_k = 2000$

тости сохраняется и особо проявляется при больших значениях k ($k > 10$).

При значениях параметра t закона распределения $p_h(c)$ больше четырех оценки ошибок модифицированного (10), традиционного непараметрического алгоритмов распознавания образов и их коллективов достоверно не отличаются при $k \in [2; 20]$.

Преимущество коллектива алгоритмов (12) со случайными коэффициентами размытости перед коллективом традиционных непараметрических классификаторов отмечается при всех $t \in [1; 7]$ и $k \in [2; 20]$. Его эффективность может быть обоснована использованием переменных ядерных мер близости между точками в пространстве признаков классифицируемых объектов, которые определяются законом распределения коэффициентов размытости. Применение принципов коллективного оценивания позволяет придать переменным мерам близости, свойственным модифицированным непараметрическим алгоритмам со случайными коэффициентами размытости, более устойчивый характер.

Заключение. Предложен рандомизированный метод оптимизации непараметрических алгоритмов распознавания образов, основанный на выборе параметров закона распределения случайных значений коэффициентов размытости ядерных функций из условия минимума эмпирической ошибки классификации. Получаемые при этом решающие правила обладают значительной устойчивостью к погрешностям определения их оптимальных параметров.

Применение коллектива непараметрических алгоритмов, соответствующих различным последовательностям случайных значений коэффициентов размытости, обеспечивает достоверное снижение ошибки распознавания образов по сравнению с традиционными непараметрическими классификаторами. Перспективность данного направления исследований состоит в возможности создания алгоритмических средств доверительного оценивания непараметрического решающего правила и его коэффициентов размытости.

СПИСОК ЛИТЕРАТУРЫ

1. **Деврой Л., Дьерди Л.** Непараметрическое оценивание плотности (L_1-подход). М.: Мир, 1988.
2. **Parzen E.** On the estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**. P. 1065.
3. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. **14**, вып. 1. С. 156.
4. **Надарая Э. А.** Об оценке плотностей распределения случайных величин // Сообщ. АН ГССР. 1964. **32**. С. 277.
5. **Бусленко Н. П., Шрейдер Ю. А.** Метод статистических испытаний. М.: Гос. изд-во физ.-мат. лит., 1961.
6. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000.
7. **Растригин Л. А.** Гибридное распознавание // АиТ. 1993. № 4. С. 3.

Поступила в редакцию 22 апреля 2006 г.