УДК 004.93

МЕТОДИКА БЫСТРОГО ВЫБОРА КОЭФФИЦИЕНТОВ РАЗМЫТОСТИ В НЕПАРАМЕТРИЧЕСКОМ КЛАССИФИКАТОРЕ, СООТВЕТСТВУЮЩЕМ КРИТЕРИЮ МАКСИМУМА АПОСТЕРИОРНОЙ ВЕРОЯТНОСТИ

© **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 университет науки и технологий*
им. академика М. Ф. Решетнева,
660037, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Предлагается быстрый алгоритм выбора коэффициентов размытости ядерных функций в непараметрическом алгоритме распознавания образов, соответствующем критерию максимума апостериорной вероятности. Основу алгоритма составляют результаты исследования асимптотических свойств непараметрической оценки уравнения разделяющей поверхности и плотностей вероятностей при решении двухальтернативной задачи распознавания образов. Проводится сравнение предложенной методики и традиционного подхода, основанного на минимизации оценки вероятности ошибки классификации.

Ключевые слова: непараметрический алгоритм распознавания образов, максимум апостериорной вероятности, ядерная оценка плотности вероятности, быстрый выбор коэффициентов размытости, оценка плотности вероятности типа Розенблатта — Парзена.

DOI: 10.15372/AUT20190610

Введение. Непараметрические алгоритмы распознавания образов, синтез которых основан на оценках плотности вероятности типа Розенблатта — Парзена, широко используются при исследовании объектов различной природы в условиях априорной неопределённости. Их вычислительная эффективность значительно снижается с увеличением объёма обучающих выборок, что особенно характерно при обработке больших объёмов данных дистанционного зондирования [1, 2]. Поэтому возникает необходимость разработки новых методик, обеспечивающих сокращение временны́х затрат при синтезе и применении непараметрических алгоритмов распознавания образов ядерного типа.

В последнее время этой проблеме уделяется значительное внимание исследователей. Предложен ряд методик быстрого выбора коэффициентов размытости ядерных функций для непараметрической оценки плотности вероятности [3–9]. Рассматриваемые процедуры основаны на результатах анализа её асимптотических свойств.

В данной работе полученные результаты используются при обосновании и разработке методики быстрого выбора коэффициентов размытости ядерных функций в непараметрическом алгоритме распознавания образов, соответствующем критерию максимума апостериорной вероятности при решении двухальтернативной задачи классификации.

Непараметрическая оценка решающей функции и её свойства. Байесовское решающее правило для двухальтернативной задачи распознавания образов, соответствующее критерию максимума апостериорной вероятности, имеет вид [10]

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

где

$$f_{12}(x) = P_2 p_2(x) - P_1 p_1(x) \tag{1}$$

— уравнение разделяющей поверхности между классами Ω_1, Ω_2; $p_j(x)$ — условная плотность вероятности распределения признаков x анализируемых объектов в классе Ω_j, $j = 1, 2$, а P_j — их априорная вероятность. Далее для простоты изложения под x понимается одномерная случайная величина.

Пусть $V = (x^i, \delta(i), i = \overline{1, n})$ — обучающая выборка объёма n, составленная из значений признака x^i классифицируемых объектов и соответствующих им «указаний учителя» $\delta(i)$ о принадлежности объектов к одному из двух классов Ω_1, Ω_2. В условиях априорной неопределённости вида плотности вероятности $p_j(x)$ в качестве плотности оценки будем использовать статистику Розенблатта — Парзена [11]

$$\bar{p_j}(x) = (n_j cj)^{-1} \sum_{i \in I_j} \Phi\left(\frac{x - x^i}{c_j}\right), \quad j = 1, 2, \tag{2}$$

где I_j — множество номеров наблюдений случайной величины x из класса Ω_j в обучающей выборке V; n_j — количество элементов множества I_j. Оценку априорной вероятности P_j класса Ω_j будем вычислять как частоту $\bar{P}_j = n_j/n$, $j = 1, 2$.

Ядерные функции $\Phi(u)$ в статистике (2) удовлетворяют следующим условиям:

$$\Phi(u) = \Phi(-u), \quad 0 \leqslant \Phi(u) < \infty, \quad \int \Phi(u) du = 1,$$

$$\int u^2 \Phi(u) du = 1, \int u^m \Phi(u) du < \infty, \quad 0 \leqslant m < \infty.$$

Здесь и далее бесконечные пределы интегрирования опускаются. Последовательности коэффициентов размытости ядерных функций $c_j = c(n_j) \to 0$, а $n_j c_j \to \infty$ при $n_j \to \infty$, $j = 1, 2$.

С учётом выражения (2) непараметрическую оценку уравнения разделяющей поверхности (1) запишем в виде

$$\bar{f}_{12}(x) = \frac{1}{nc} \sum_{i=1}^{n} \delta(i) \Phi\left(\frac{x - x^i}{c}\right), \tag{3}$$

где

$$\delta(i) = \begin{cases} -1, & \text{если } x^i \in \Omega_1; \\ 1, & \text{если } x^i \in \Omega_2, \end{cases}$$

а c — коэффициент размытости ядерных функций в статистике (3). Проведём анализ среднеквадратического отклонения $\bar{f}_{12}(x)$ от $f_{12}(x)$ при достаточно больших объёмах обучающей выборки

$$W = M \int \left(\bar{f}_{12}(x) - f_{12}(x)\right)^2 dx = M \int \left(P_1 p_1(x) - \bar{P}_1 \bar{p}_1(x)\right)^2 dx -$$

$$-2M\int\big(P_1p_1(x)-\bar{P}_1\bar{p}_1(x)\big)\big(P_2p_2(x)-\bar{P}_2\bar{p}_2(x)\big)\,dx+M\int\big(P_2p_2(x)-\bar{P}_2\bar{p}_2(x)\big)^2\,dx, \quad (4)$$

где M — знак математического ожидания.

Рассмотрим два подхода к оптимизации непараметрической оценки уравнения разделяющей поверхности (3) по коэффициенту размытости ядерной функции c. В первом случае оптимальное значение c^* коэффициента размытости ядерных функций определяется из условия минимума асимптотического выражения $\bar{W}(c)$ критерия (4). Будем использовать технологию доказательства асимптотических свойств непараметрических оценок плотностей вероятности Епанечникова [12]. Тогда при достаточно больших значениях n_j, $j=1,2$, среднеквадратическое отклонение (4) определяется выражением [13]

$$\bar{W}(c)=\frac{\|\Phi(u)\|^2}{c}\Big(\frac{P_2^2}{n_2}+\frac{P_1^2}{n_1}\Big)+\frac{c^4}{4}B, \quad (5)$$

где $B=\int(P_2p_2^{(2)}(x)-P_1p_1^{(2)}(x))^2dx$, $\|\Phi(u)\|^2=\int\Phi^2(u)du$, $p_j^{(2)}(x)$ — вторая производная плотности вероятности $p_j(x)$, $j=1,2$.

В этом случае оптимальный коэффициент размытости c^*, минимизирующий выражение (5), соответствует значению

$$c^*=\Big(\frac{\|\Phi(u)\|^2(P_2^2n_1+P_1^2n_2)}{n_1n_2B}\Big)^{1/5}. \quad (6)$$

Тогда минимальное значение асимптотического выражения среднеквадратического отклонения $\bar{f}_{12}(x)$ от $f_{12}(x)$ (5) при $c=c^*$ запишем в виде

$$\bar{W}(c^*)=\frac{5}{4}\Big[\Big(\frac{\|\Phi(u)\|^2(P_2^2n_1+P_1^2n_2)}{n_1n_2}\Big)^4B\Big]^{1/5}. \quad (7)$$

Второй подход к оптимизации $\bar{f}_{12}(x)$ состоит в выборе коэффициентов размытости ядерных функций c_1, c_2 из условия минимума асимптотических выражений среднеквадратических отклонений $\bar{p}_j(x)$ от $p_j(x)$ [12]:

$$\bar{W}_j(c_j)\sim\frac{\|\Phi(u)\|^2}{n_jc_j}+\frac{c_j^4B_j}{4},\quad j=1,2, \quad (8)$$

которые соответствуют составляющим $M\int(p_j(x)-\bar{p}_j(x))^2dx$ критерия (4). В выражении (8) значения $B_j=\int(p_j^{(2)}(x))^2dx$.

Из условия минимума выражения (8) по c_j получим

$$c_j^*=\Big[\frac{\|\Phi(u)\|^2}{n_jB_j}\Big]^{1/5},\quad j=1,2, \quad (9)$$

а соответствующий им критерий (4) определяется выражением

$$\bar{W}(c_1^*,c_2^*)\sim\frac{5(\|\Phi(u)\|^2)^{4/5}}{4}\Big[P_1^2\Big(\frac{B_1}{n_1^4}\Big)^{1/5}+P_2^2\Big(\frac{B_2}{n_2^4}\Big)^{1/5}\Big]-\frac{P_1P_2}{2}B_{12}\Big(\frac{\|\Phi(u)\|^2}{\prod\limits_{j=1}^{2}n_jB_j}\Big)^{2/5}. \quad (10)$$

Таблица 1

Зависимости отношения $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*)$ от значений σ_2, n_2
при математических ожиданиях $m_1 = 0$, $m_2 = 3$ случайных величин x в классах Ω_1, Ω_2 и $\sigma_1 = 1$

Значения σ_2	Значения n_2									
	100	200	300	400	500	600	700	800	900	1000
0,25	1,4	1,21	1,18	1,14	1,11	1,1	1,09	1,08	1,07	1,05
0,5	1,19	1,16	1,12	1,11	1,09	1,07	1,07	1,04	1,04	1,06
1	0,98	0,99	0,99	1	1	1	1	1	1	1
2	1,07	1,07	1,06	1,03	1,02	1,01	1,01	1	1	0,99
4	1,4	1,54	1,57	1,57	1,55	1,52	1,49	1,46	1,43	1,4

Здесь введено обозначение

$$B_{12} = \int p_1^{(2)}(x) p_2^{(2)}(x) dx.$$

Проведём сравнение критериев (7), (10) оценок среднеквадратических ошибок аппроксимации статистикой $\bar{f}_{12}(x)$ (3) байесовского уравнения разделяющей поверхности $f_{12}(x)$ (1) с позиций двух подходов к оптимизации $\bar{f}_{12}(x)$ по коэффициентам размытости ядерных функций. Будем считать, что законы распределения случайных величин в классах Ω_1, Ω_2 подчиняются законам Гаусса $N_1(0;1)$, $N_2(3;\sigma_2)$. Значения среднеквадратического отклонения σ_2 и n_2 изменяются в процессе вычислительных экспериментов при $n_1 = 100$. Полученные результаты представлены в табл. 1, элементы которой соответствуют значениям отношения $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*)$.

С ростом объёма обучающей выборки $n = n_1 + n_2$ значения критериев $\bar{W}(c^*)$, $\bar{W}(c_1^*, c_2^*)$ уменьшаются при изменении среднеквадратического отклонения σ_2, что согласуется с результатами исследований асимптотических свойств непараметрических статистик $\bar{f}_{12}(x)$, $\bar{p}_j(x)$, $j = 1, 2$. Например, при $\sigma_2 = 1$ с ростом n_2 в интервале [100; 1000] значения $\bar{W}(c^*)$ и $\bar{W}(c_1^*, c_2^*)$ убывают в интервале [0,004; 0,001]. При $\sigma_2 = 2$ в условиях $n_2 \in [100; 1000]$ значения $\bar{W}(c^*)$ снижаются от значения 0,0036 до 0,0006, а критерий $\bar{W}(c_1^*, c_2^*)$ убывает от 0,0033 до значения 0,0006.

Из анализа результатов вычислительных экспериментов следует, что выбор коэффициентов размытости ядерных функций из условия минимума среднеквадратического отклонения $\bar{p}_j(x)$ от $p_j(x)$, $j = 1, 2$, типа (8) способствует более низкому уровню ошибки аппроксимации статистикой $\bar{f}_{12}(x)$ решающей функции $f_{12}(x)$. Этот вывод следует из выполнения условий $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*) > 1$, представленных в табл. 1. Отмеченная тенденция проявляется особенно значимо при увеличении априорной вероятности Ω_2 в условиях $\sigma_2 > \sigma_1$, когда плотности вероятности $p_1(x)$, $p_2(x)$ отличаются значительно.

Обнаружен экстремальный характер зависимостей отношений $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*)$ от значений σ_2 при фиксированных значениях σ_1 и объёме выборки n. С ростом априорной вероятности второго класса этот экстремум незначительно смещается в интервал больших значений σ_2 (рисунок).

Результаты анализа данных вычислительных экспериментов показывают, что значения критериев (7), (10) в достаточно широких условиях эксперимента сопоставимы. Отсюда следует возможность выбора коэффициентов размытости ядерных функций в непараметрической оценке уравнения разделяющей поверхности (3) из условия минимума среднеквадратических критериев $\bar{W}_j(c_j)$, $j = 1, 2$, ошибок аппроксимации плотностей вероятностей $p_j(x)$ статистиками $\bar{p}_j(x)$, $j = 1, 2$, типа (2).

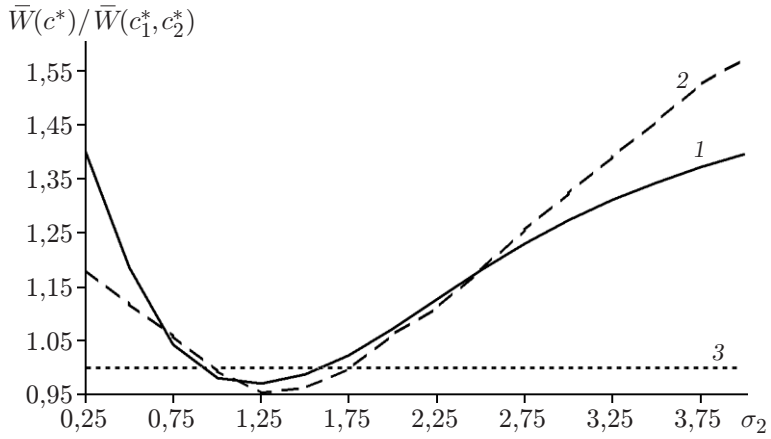

Иллюстрация зависимостей отношения $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*)$ от значений σ_2, n_2. Кривая *1* соответствует $n_1 = n_2 = 100$, а кривая *2* — значениям $n_1 = 100$ и $n_2 = 300$. Прямая *3* определяет значение $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*) = 1$

Методика быстрого выбора коэффициентов размытости ядерной оценки уравнения разделяющей поверхности. Следуя рекомендациям [8, 9], представим выражение (9) в виде произведения:

$$c_j^* = \beta_j \sigma_j n_j^{-1/5}, \quad j = 1, 2, \tag{11}$$

где

$$\beta_j = (\|\Phi(u)\|^2/(B_j \sigma_j^5))^{1/5}.$$

Формулы (9), (11) являются тождественными. Значения β_j при выбранной ядерной функции определяются только видом плотности вероятности и не зависят от её параметров. Например, для нормального, логистического, Лапласа и экспоненциального законов распределения параметр β_j принимает соответственно значения 1,049; 0,895; 0,717; 0,883 при использовании ядерной функции Епанечникова (табл. 2). Приведённые значения β_j являются фрагментами тестового семейства плотностей вероятности [8].

Таблица 2

Зависимости значений параметра β_j
от законов распределения случайных величин и вида ядерных функций

Закон распределения	Вид ядерных функций		
	ступенчатое ядро	ядро Епанечникова	гауссовское ядро
Нормальный	1,064	1,049	1,059
Логистический	0,908	0,895	0,904
Лапласа	0,728	0,717	0,724
Экспоненциальный	0,896	0,883	0,892
Односторонний нормальный	0,926	0,913	0,922
Односторонний логистический	1,226	1,208	1,22

В условиях неопределённости вида восстанавливаемой плотности вероятности предложена оценка коэффициента размытости ядерных функций [8]

$$\bar{c}_j^* = \bar{\beta}_j \bar{\sigma}_j n_j^{-1/5}, \quad j = 1, 2. \tag{12}$$

Значения $\bar{\beta}_j = N^{-1} \sum\limits_{t=1}^{N} \beta(t)$ вычисляются по значениям параметров

$$\beta(t) = (\|\Phi(u)\|^2 / (B(t)\sigma^5(t)))^{1/5}, \quad t = \overline{1, N},$$

тестового семейства плотностей вероятностей в количестве N. В табл. 2 значение $N = 6$. Тогда методика быстрого выбора коэффициентов размытости ядерных функций непараметрического уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов типа (3) сводится к выполнению следующих действий:

1. По обучающей выборке V сформировать последовательности $V_j = (x^i, i \in I_j)$, $j = 1, 2$.

2. Определить оценки среднеквадратических отклонений

$$\bar{\sigma}_j = \Big(\frac{1}{n_j - 1} \sum_{i \in I_j} (x^i - \bar{x}_j)^2 \Big)^{1/2}, \quad j = 1, 2,$$

где n_j — количество элементов множества I_j, а $\bar{x}_j$ — среднее значение случайной величины x в классе Ω_j.

3. Используя значения параметров $\beta(t)$, $t = \overline{1, N}$, тестового семейства плотностей вероятности, соответствующих выбранному виду ядерной функции, определить $\bar{\beta}_j$, $j = 1, 2$, как их среднее значение.

4. При известных значениях n_j и $\bar{\beta}_j$ вычислить по формуле (12) оценки $\bar{c}_j^*$, $j = 1, 2$, оптимальных коэффициентов размытости ядерных функций для непараметрических оценок плотностей вероятности типа (2).

5. Построить по выборкам $V_j = (x^i, i \in I_j)$, $j = 1, 2$, непараметрические оценки плотностей вероятностей (2) при $c_j = \bar{c}_j^*$, $j = 1, 2$.

6. На основании результатов этапа 5 осуществить синтез непараметрической оценки уравнения разделяющей поверхности (3), которая является базой построения оценки решающего правила распознавания образов $\bar{m}(x)$.

В качестве примера использования предлагаемой методики быстрого выбора коэффициента размытости ядерных функций в статистике (3) будем решать двухальтернативную задачу распознавания образов по обучающей выборке $V = (x^i, \delta(i), i = \overline{1, n})$. Выборка V формируется с помощью датчиков случайных величин с нормальными законами распределения $p_1(x) = N_1(0; 1)$, $p_2(x) = N_2(3, 2)$, где значения 0 и 3 соответствуют их математическим ожиданиям в классах Ω_j, $j = 1, 2$, а 1 и 2 — среднеквадратическим отклонениям.

Пусть количество ситуаций первого класса в выборке V определяется значением $n_1 = 100$, а второго — $n_2 = 500$. В данных условиях вычислим оценки среднеквадратических отклонений $\bar{\sigma}_1$, $\bar{\sigma}_2$ случайной величины x в классах. Они соответствуют значениям $\bar{\sigma}_1 = 0,887$, $\bar{\sigma}_2 = 2,043$. Определим влияние погрешностей их оценивания на значения оценок вероятностей ошибок распознавания образов. Для этого построим доверительные интервалы для σ_1, σ_2 при коэффициенте доверия $\beta = 0,95$ [14]. При заданных значениях β и степенях свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ установим доверительные интервалы для σ_1, σ_2: $0,757 < \sigma_1 < 1,017$; $1,912 < \sigma_2 < 2,174$. Обозначим через $\bar{\sigma}_j(1)$, $\bar{\sigma}_j(2)$ левую и правую доверительные границы для σ_j в классах Ω_j, $j = 1, 2$.

Таблица 3

Оценки вероятностей ошибки распознавания образов
при различных значениях коэффициентов размытости статистик $\bar{p}_1(x)$, $\bar{p}_2(x)$

Значения коэффициентов размытости	$\bar{c}_1^*(1)$	$\bar{c}_1^*(2)$	$\bar{c}_1^*$
$\bar{c}_2^*(1)$	0,14	0,137	0,14
$\bar{c}_2^*(2)$	0,14	0,138	0,137
$\bar{c}_2^*$	0,14	0,137	0,138

При реализации последующих этапов рассматриваемой методики в процессе синтеза непараметрических оценок плотностей вероятностей и решающего правила $\bar{m}(x)$ будем использовать ядерные функции Епанечникова [12]

$$\Phi(u) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3}{20\sqrt{5}}u^2, & \text{если } |u| < \sqrt{5}, \\ 0, & \text{если } |u| \geqslant \sqrt{5}. \end{cases}$$

В этих условиях $\|\Phi(u)\|^2 = 3/(5\sqrt{5})$.

По информации, содержащейся в табл. 2, вычислим значения параметров $\bar{\beta}_1 = \bar{\beta}_2 = 0,9232$ формулы (12). Так как обучающая выборка формируется с использованием нормальных законов распределения случайной величины x в классах, то в табл. 2 соответствующие им данные исключаются.

По формуле (12) вычислим последовательность оценок оптимальных коэффициентов размытости ядерных функций для статистик типа (2) с учётом погрешности определения по исходным данным значений среднеквадратических отклонений в классах:

$$\bar{c}_1^*(1) = 0,278; \quad \bar{c}_1^*(2) = 0,374; \quad \bar{c}_1^* = 0,326;$$

$$\bar{c}_2^*(1) = 0,509; \quad \bar{c}_2^*(2) = 0,579; \quad \bar{c}_2^* = 0,544.$$

Здесь $\bar{c}_j^*(1)$, $\bar{c}_j^*(2)$, $\bar{c}_j^*$ соответствуют значениям $\bar{\sigma}_j(1)$, $\bar{\sigma}_j(2)$, $\bar{\sigma}_j$, $j = 1, 2$.

Результаты решения задачи распознавания образов в условиях рассматриваемого примера при различных значениях коэффициентов размытости непараметрических оценок плотностей вероятностей в классах приведены в табл. 3.

Оценки вероятностей ошибки распознавания образов с использованием непараметрического решающего правила $\bar{m}(x)$ вычислялись в режиме «скользящего экзамена», например, в соответствии с выражением

$$\bar{\rho}(\bar{c}_1^*(1), \bar{c}_2^*(1)) = \frac{1}{n}\sum_{j=1}^{n} 1(\delta(j), \bar{\delta}(j)), \tag{13}$$

где

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

$\bar{\delta}(j)$ — «решение» о принадлежности ситуации x^j к одному из двух классов с использованием статистики (3) в правиле $\bar{m}(x)$; $\delta(j)$ — «указания» из обучающей выборки V относительно принадлежности x^j к Ω_1 либо Ω_2. Если ситуация x^j из обучающей выборки V подаётся на контроль в выражение (13), то она исключается из процесса обучения

в статистике (3). В этих условиях принятие решений алгоритмом $\bar{m}(x)$ осуществляется на основе непараметрической оценки уравнения разделяющей поверхности $\bar{f}_{12}(x)$ (3) при $x^i \neq x^j$. Для традиционного подхода оценка оптимального коэффициента размытости $\bar{c}^*$ ядерных функций статистики (3) определяется из условия минимума критерия типа (13). Его минимальное значение в принятых условиях вычислительных экспериментов равно 0,135 при $\bar{c}^* = 0,65$.

Достоверность сравнения результатов табл. 3 со значением $\bar{\rho}(\bar{c}^*) = 0,135$ проверялась в соответствии с методикой Колмогорова — Смирнова [15] и критерием проверки гипотезы о равенстве вероятностей [14]. Пороговое значение критерия Колмогорова — Смирнова

$$D_\alpha = \sqrt{-\ln\frac{\alpha}{2}(\frac{1}{n_1}+\frac{1}{n_2})/2}$$

в рассматриваемых условиях вычислительных экспериментов равно 0,149 при риске $\alpha = 0,05$ отвергнуть, например, гипотезу $H_0 : \rho(c_1^*(1), c_2^*(1)) = \rho(c^*)$. Здесь $\rho(c^*)$, $\rho(c_1^*(1), c_2^*(1))$ — вероятности ошибок распознавания образов, оценки которых вычисляются по формуле типа (13), а c^*, c_1^*, c_2^* — оптимальные коэффициенты размытости ядерных функций статистики (3) и непараметрических оценок плотностей вероятностей вида (2). Нетрудно заметить, что $\bar{D} = |\rho(c_1^*(1), c_2^*(1)) - \rho(c^*)| = 0,003$ намного меньше значения D_α. Данный вывод стправедлив и для других условий табл. 3 и подтверждается результатами использования традиционного критерия о равенстве вероятностей [14]. Поэтому предлагаемая методика выбора коэффициентов размытости ядерных функций в непараметрической оценке уравнения разделяющей поверхности (3) позволяет получать значения вероятности ошибки распознавания образов, сопоставимые с традиционным подходом. Её применение позволяет на порядки сократить временны́е затраты при выборе оценки оптимального коэффициента размытости ядерных функций. Высокая вычислительная эффективность предлагаемой методики объясняется исключением процедуры поиска экстремума оценки вероятности ошибки распознавания образов, что свойственно традиционному подходу.

Исследуем оценки вероятностей ошибок распознавания образов при использовании предлагаемой методики выбора коэффициента размытости ядерной функции в статистике (3) и традиционного подхода оптимизации непараметрического алгоритма классификации. Условия вычислительных экспериментов отличаются параметрами плотности вероятности распределения случайных величин во втором классе и объёмом обучающей выборки, которая формируется по нормальным законам распределения $p_1(x) = N_1(0;1)$, $p_2(x) = N_2(3;\sigma_2)$.

Результаты анализа данных вычислительных экспериментов представлены в табл. 4, 5. Элементы таблиц соответствуют отношениям $w = \bar{\rho}(\bar{c}^*)/\bar{\rho}(\bar{c}_1^*, \bar{c}_2^*)$ оценок вероятностей ошибок распознавания образов, вычисляемых с использованием традиционной и предлагаемой методик выбора коэффициентов размытости ядерных функций в непараметрической оценке уравнения разделяющей поверхности (3). Значения w определялись как средние по 50 результатам вычислительных экспериментов. По обучающей выборке осуществляется синтез непараметрического алгоритма распознавания образов $\bar{m}(x)$, а контрольная выборка формируется в соответствии с условиями вычислительного эксперимента.

Из анализа информации вычислительных экспериментов следует сопоставимость оценок вероятностей ошибок распознавания образов $\rho(c^*)$, $\rho(c_1^*, c_2^*)$ при использовании предлагаемого и традиционного методов оптимизации непараметрического классификатора (см. табл. 4, 5). Для подтверждения данного факта проверим достоверность отличия оценок вероятностей ошибок распознавания образов. В условиях табл. 4 пороговое значение критерия Колмогорова — Смирнова D_α принадлежит интервалу [0,118; 0,104] при

Таблица 4

Зависимости w от объёма выборки n_2 и значений σ_2.
Оценки вероятностей ошибок распознавания образов определяются
по контрольной выборке объёма $n_k = 200 + 2n_2$ ($n_{k1} = 200$, $n_{k2} = 2n_2$)

Значения σ_2	Значения n_2				
	200	300	400	500	600
0,25	1,67	1,55	1,76	1,79	1,75
0,5	1,14	1,20	1,11	1,17	1,16
1	1,04	1,04	1,04	1,03	1,02
2	1,02	1,05	1,03	1,03	1,02
3	1,05	1,06	1,06	1,02	1,03
4	1,03	1,05	1,06	1,03	1,03

Таблица 5

Зависимости w от объёма выборки n_2 и значений σ_2.
Оценки вероятностей ошибок распознавания образов определяются
по обучающей выборке объёма $n = (n_1 + n_2)$ в условиях $n_1 = 100$

Значения σ_2	Значения n_2				
	200	300	400	500	600
0,25	1,23	1,37	1,22	1,35	1,14
0,5	0,98	0,97	0,98	0,99	1,03
1	0,94	0,97	0,95	0,95	0,96
2	0,98	0,98	0,98	0,99	0,98
3	0,99	1,00	1,01	1,00	0,99
4	1,00	1,00	1,01	0,99	0,99

изменении значений объёма контрольной выборки $n_k \in [600;\ 1400]$. При этом максимальное значение $\bar{D}$ не превышает величины 0,1 в процессе анализа каждого условия (n_2, σ_2) табл. 4 во всех 50 вычислительных экспериментах, что обосновывает сопоставимость исследуемых методик. Аналогичные результаты и выводы получены при анализе данных табл. 5. В этом случае $D_\alpha \in [0,166;\ 0,147]$ при $n \in [300;\ 700]$.

Сопоставимость и достоверность результатов вычислительных экспериментов, приведённых в табл. 4 и 5, подтверждается также при использовании традиционного подхода сравнения двух вероятностей ошибок распознавания образов $\rho(c^*)$, $\rho(c_1^*, c_2^*)$ [14]. Установлено, что максимальное наблюдаемое значение критерия $u_{\text{н}}$ в принятых условиях вычислительных экспериментов в несколько раз меньше критического $u_{\text{кр}} = 1{,}95$, которое определялось по таблице функции Лапласа при уровне значимости $\alpha = 0{,}05$.

Однако при оценивании значений $\bar{\rho}(\bar{c}^*)$, $\bar{\rho}(\bar{c}_1^*, \bar{c}_2^*)$ по контрольной выборке наблюдается преимущество предлагаемой методики по сравнению с традиционной при всех $\sigma_2 \in [0,25; 4]$, $n_2 \in [200; 600]$ (см. табл. 4). Этот факт особенно проявляется при малых значениях σ_2. Тогда при сложных решающих функциях, содержащих оценки априорных вероятностей классов и соответствующих им плотностей вероятности, особое внимание следует уделять точности аппроксимации статистиками $\bar{p}_j(x)$ функций $p_j(x)$, $j = 1, 2$.

При оценивании вероятности ошибки распознавания образов по обучающей выборке с использованием предлагаемой методики выбора коэффициентов размытости ядерных функций наблюдается незначительное снижение её эффективности при $\sigma_2 > 0,25$ по сравнению с традиционным подходом (см. табл. 5). Данный факт объясняется особенностью метода «скользящего экзамена», когда определение коэффициентов размытости в стати-

стике (3) ориентировано на минимизацию критерия (13). Естественно, что данные условия обеспечивают незначительное преимущество традиционного подхода.

Результаты вычислительных экспериментов согласуются с выводами асимптотических исследований. Значения коэффициентов размытости ядерных функций непараметрической оценки плотности вероятности $\bar{p}_1(x)$ соответствуют значению $\bar{c}_1^* = 0,41$ и не изменяются в процессе вычислительных экспериментов, так как $n_1 = 100$. С ростом n_2 в интервале [200; 600] значения $\bar{c}_2^*$ уменьшаются от 0,18 до 0,14 при $\sigma_2 = 0,5$. При $\sigma_2 = 1$ значение $\bar{c}_2^* \in [0,36;\ 0,22]$ для приведённых выше условий.

Увеличение области определения $p_2(x)$ естественно сопровождается повышением значений оптимального коэффициента размытости ядерной функции $\bar{c}_2^*$. Например, с ростом σ_2 и n_2 в интервале [200; 600] значение $\bar{c}_2^* \in [0,72;\ 0,58]$. С увеличением интервала пересечения классов наблюдается рост вероятности ошибки распознавания образов. При $\sigma_2 = 1$ и $n_2 \in [200;\ 600]$ значения $\bar{\rho}(\bar{c}^*)$, вычисляемые по контрольной выборке $n_{k1} = 200$, $n_{k2} = 400$, принимают значения в интервале [0,065; 0,043]. В этих условиях при $\sigma_2 = 2$ значения $\bar{\rho}(\bar{c}^*) \in [0,16;\ 0,13]$.

Заключение. Традиционный подход к оптимизации непараметрических алгоритмов распознавания образов, соответствующих критерию максимума апостериорной вероятности, предполагает выбор коэффициентов размытости ядерных функций из условия минимума оценки вероятности ошибки классификации в режиме «скользящего экзамена». Его вычислительная эффективность значительно снижается с увеличением объёма обучающей выборки. На основе анализа асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов установлена возможность выбора коэффициентов размытости ядерных функций по результатам оптимизации непараметрических оценок плотностей вероятностей случайных величин в классах. Этот вывод позволяет методику быстрого выбора коэффициентов размытости для ядерных оценок плотностей вероятностей развить на решение задачи оптимизации непараметрической оценки уравнения разделяющей поверхности между классами и на порядки сократить временны́е затраты на её решение. По результатам вычислительных экспериментов установлено превышение значений оценки вероятности ошибки распознавания образов для традиционного подхода оптимизации решающего правила классификации над соответствующим показателем эффективности при использовании предлагаемой методики выбора коэффициентов размытости. Данный вывод сохраняется при достаточно широких условиях вычислительных экспериментов, связанных с изменением объёма обучающей выборки и интервалом пересечения классов. Полученные результаты обеспечивают их развитие при выборе коэффициентов размытости ядерных функций для многомерной непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов.

Финансирование. Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 18-01-00251).

СПИСОК ЛИТЕРАТУРЫ

1. **Дмитриев Е. В., Козодеров В. В., Дементьев А. О., Сафонова А. Н.** Комплексирование классификаторов в задаче тематической обработки гиперспектральных аэрокосмических изображений // Автометрия. 2018. **54**, № 3. С. 3–13. DOI: 10.15372/AUT20180301.
2. **Лапко А. В., Лапко В. А.** Непараметрические алгоритмы оценивания состояний природных объектов // Автометрия. 2018. **54**, № 5. С. 33–39. DOI: 10.15372/AUT20180504.
3. **Borrajo M. I., González-Manteiga W., Martínez-Miranda M. D.** Bandwidth selection for kernel density estimation with length-biased data // Journ. Nonparametr. Statist. 2017. **29**, Iss. 3. P. 636–668. DOI: 10.1080/10485252.2017.1339309.

4. **Chen S.** Optimal bandwidth selection for kernel density functionals estimation // Journ. Probability and Statist. 2015. **242683**. DOI: 10.1155/2015/242683.
5. **Scott D. W.** Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley and Sons, 2015. 384 p.
6. **Sheather S. J.** Density estimation // Statist. Sci. 2004. **19**, N 4. P. 588–597. DOI: 10.1214/088342304000000297.
7. **Silverman B. W.** Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986. 175 p.
8. **Лапко А. В., Лапко В. А.** Быстрый алгоритм выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности // Измерительная техника. 2018. № 6. С. 16–20. DOI: 10.32446/0368-1025it-2018-6-16-20.
9. **Лапко А. В., Лапко В. А.** Быстрый алгоритм выбора коэффициентов размытости в многомерных ядерных оценках плотности вероятности // Измерительная техника. 2018. № 10. С. 19–23.
10. **Лапко А. В., Лапко В. А.** Многоуровневые непараметрические системы обработки информации. Красноярск: СибГАУ, 2013. 270 с.
11. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
12. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. **14**, № 1. С. 156–161.
13. **Лапко А. В., Лапко В. А.** Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности в алгоритме распознавания образов, соответствующего критерию максимума апостериорной вероятности // Системы управления и информационные технологии. 2010. **42**, № 4. С. 58–61.
14. **Гмурман В. Е.** Теория вероятностей и математическая статистика. М.: Высш. шк., 1999. 479 с.
15. **Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А.** Сложные системы. М.: Высш. шк., 1977. 248 с.

Поступила в редакцию 09.07.2018
После доработки 19.10.2018
Принята к публикации 22.10.2018