УДК 681.513

СИНТЕЗ СТРУКТУРЫ СЕМЕЙСТВА НЕПАРАМЕТРИЧЕСКИХ РЕШАЮЩИХ ФУНКЦИЙ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ*

А. В. Лапко[1], **В. А. Лапко**[1, 2]

[1] *Учреждение Российской академии наук*
Институт вычислительного моделирования Сибирского отделения РАН,
660036, г. Красноярск, Академгородок, 50, стр. 44
[2] *Государственное образовательное учреждение высшего профессионального образования*
«Сибирский государственный аэрокосмический университет
им. академика М. Ф. Решетнёва»,
660014, г. Красноярск, просп. «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

На основе анализа асимптотических свойств семейства непараметрических решающих функций в задаче распознавания образов предлагается методика синтеза его структуры.

Ключевые слова: семейство решающих функций, распознавание образов, непараметрическая оценка, большие выборки, структура.

Введение. Непараметрические алгоритмы распознавания образов, основанные на оценках плотности вероятности типа Розенблатта — Парзена, активно используются при исследовании объектов различной природы в условиях априорной неопределённости. Однако их вычислительная эффективность значительно снижается с ростом объёма статистических данных, так как формирование каждого решения требует анализа всех элементов.

В работе [1] обосновано направление «обхода» возникающей проблемы при оценивании решающей функции в задаче распознавания образов для условий больших выборок с использованием принципа декомпозиции и технологии параллельных вычислений. При этом разработаны двухуровневые непараметрические системы для решения двуальтернативной [2] и многоальтернативной [1] задач классификации, установлены асимптотические свойства их уравнений разделяющих поверхностей для одномерного случая. Показано, что исследуемая статистика по сравнению с традиционной непараметрической оценкой решающей функции типа Розенблатта — Парзена имеет меньшую дисперсию, а сокращение времени вычисления её значений сопоставимо с количеством составляющих семейства.

Цель данной работы состоит в создании на основе анализа аппроксимационных свойств семейства многомерных непараметрических решающих функций методики синтеза и анализа его рациональной структуры.

Семейство многомерных непараметрических решающих функций и его свойства. Имеется обучающая выборка $V = (x^i,\, \sigma(x^i),\, i = \overline{1,n})$ объёма n, составленная из признаков $x^i = (x^i_j,\, j = \overline{1,k})$ классифицируемых объектов и соответствующих «указаний учителя» $\sigma(x^i)$ об их принадлежности к одному из двух классов Ω_1, Ω_2:

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

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

Обучающая выборка V обладает большим объёмом n, что снижает вычислительную эффективность традиционных непараметрических алгоритмов распознавания образов.

В соответствии с методикой, изложенной в работах [1, 2], осуществим декомпозицию исходной выборки V на части $V_j = (x^i,\, \sigma(x^i),\, i \in I_j),\, j = \overline{1, N}$, где I_j — множество номеров ситуаций из V, составляющих j-ю группу. Количество элементов n_j множеств $I_j,\, j = \overline{1, N}$, одинаково и равно $\bar{n}$, причём $N = n/\bar{n}$.

На этой основе построим семейство непараметрических оценок уравнения разделяющей поверхности

$$\bar{f}_{12}^{j}(x) = \bar{p}_2^j(x) - \bar{p}_1^j(x), \quad j = \overline{1, N}. \tag{1}$$

В выражении (1) непараметрическая оценка плотности вероятности распределения признаков x анализируемых объектов в s-м классе представляется статистикой типа Розенблатта — Парзена [3, 4]

$$\bar{p}_s^j(x) = \frac{1}{\bar{n}_s \prod\limits_{v=1}^{k} c_v} \sum_{i \in I_j^s} \prod_{v=1}^{k} \Phi\Big(\frac{x_v - x_v^i}{c_v}\Big), \quad s = \overline{1, 2},$$

где I_j^s — множество номеров ситуаций s-го класса в выборке V_j, а $\bar{n}_s$ — их количество. Ядерные функции $\Phi(u_v)$ удовлетворяют условиям H:

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

$$\int \Phi(u_v)du_v = 1, \quad \int u_v^2 \Phi(u_v)du_v = 1,$$

$$\int u_v^m \Phi(u_v)du_v < \infty, \quad 0 \le m < \infty, \;\; v = \overline{1, k},$$

где $c_v,\, v = \overline{1, k}$, — коэффициенты размытости ядерных функций, значения которых убывают с ростом количества $\bar{n}$ элементов множества $I_j,\, j = \overline{1, N}$.

Здесь и далее бесконечные пределы интегрирования опускаются.

В качестве обобщённой непараметрической решающей функции примем линейный функционал вида

$$\overline{\bar{f}}_{12}(x) = \frac{1}{N} \sum_{j=1}^{N} \bar{f}_{12}^{j}(x), \tag{2}$$

в котором количество N составляющих семейства априори задано.

Оптимизация непараметрического решающего правила

$$\bar{m}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \overline{\bar{f}}_{12}(x) \le 0, \\ x \in \Omega_2, & \text{если } \overline{\bar{f}}_{12}(x) > 0, \end{cases} \tag{3}$$

по коэффициентам размытости ядерных функций $c_v,\, v = \overline{1, k}$, осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки

распознавания образов

$$\bar{\rho}_j = \frac{1}{\bar{n}} \sum_{t \in I_j} 1(\sigma(x^t),\, \bar{\sigma}(x^t)), \quad j = \overline{1, N},$$

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

При формировании «решения» $\bar{\sigma}(x^t)$ о принадлежности ситуации x^t к одному из двух классов она исключается из процесса обучения в непараметрической статистике (2).

Без существенной потери общности будем считать, что интервалы изменения признаков x_v, $v = \overline{1, k}$, классифицируемых объектов одинаковы. В этих условиях появляется возможность полагать, что коэффициенты размытости c_v, $v = \overline{1, k}$, ядерных функций соизмеримы, т. е. $c_v = c$, $v = \overline{1, k}$.

Отметим, что статистика (2) допускает использование технологии параллельных вычислений в условиях больших выборок.

Асимптотические свойства семейства непараметрических оценок решающих функций $\overline{\overline{f}}_{12}(x)$ определяются следующим утверждением.

Теорема. Пусть плотности вероятности $p_j(x)$, $j = 1, 2$, распределения многомерной случайной величины $x = (x_1, \ldots, x_k)$ в классах и первые две их производные по каждой компоненте x_v, $v = \overline{1, k}$, ограничены и непрерывны; ядерные функции $\Phi(u)$ удовлетворяют условиям нормированности, положительности и симметричности H; последовательности $c(\bar{n}) = c$ коэффициентов размытости ядерных функций в статистиках $\bar{f}^j_{12}(x)$ таковы, что при $\bar{n}_1 \to \infty$, $\bar{n}_2 \to \infty$ значения $c \to 0$ и $\dfrac{\bar{n}_1 + \bar{n}_2}{\bar{n}_1 \bar{n}_2 c^k} \to 0$.

Тогда при конечных значениях N:

1) асимптотическое смещение семейства непараметрических решающих функций $\overline{\overline{f}}_{12}(x)$ определяется выражением

$$\mathrm{M}(f_{12}(x) - \overline{\overline{f}}_{12}(x)) \sim \frac{c^2}{2} \sum_{v=1}^{k} (p^{(2)}_{2v}(x) - p^{(2)}_{1v}(x)), \tag{4}$$

где M — знак математического ожидания, а $p^{(2)}_{jv}(x)$ — вторая производная плотности вероятности $p_j(x)$ по компоненте x_v, $v = \overline{1, k}$, $j = 1, 2$;

2) асимптотическое выражение среднеквадратического отклонения $\overline{\overline{f}}_{12}(x)$ от байесовской решающей функции $f_{12}(x) = p_2(x) - p_1(x)$ представляется в виде

$$\mathrm{M} \int \ldots \int (f_{12}(x) - \overline{\overline{f}}_{12}(x))^2 dx_1 \ldots dx_k \sim \frac{(\bar{n}_1 + \bar{n}_2) \prod\limits_{v=1}^{k} \int \Phi^2(u_v) du_v}{N \bar{n}_1 \bar{n}_2 c^k} + \frac{c^4}{4} B, \tag{5}$$

где

$$B = \int \ldots \int \Big(\sum_{v=1}^{k} (p^{(2)}_{2v}(x) - p^{(2)}_{1v}(x)) \Big)^2 dx_1 \ldots dx_k.$$

Нетрудно заметить, что при выполнении условий $c \to 0$ и $\dfrac{\bar{n}_1 + \bar{n}_2}{\bar{n}_1 \bar{n}_2 c^k} \to 0$ для $\bar{n}_s \to \infty$, $s = 1, 2$, многомерная непараметрическая оценка $\overline{\overline{f}}_{12}(x)$ сходится в среднеквадратическом к байесовскому уравнению разделяющей поверхности $f_{12}(x)$, а с учётом свойства её асимптотической несмещённости является состоятельной оценкой $f_{12}(x)$.

Доказательство утверждений теоремы об асимптотических свойствах семейства $\overline{\overline{f}}_{12}(x)$ многомерных непараметрических решающих функций основывается на технологии, предложенной в работе [1] для одномерного случая.

При оценивании байесовского уравнения разделяющей поверхности $f_{12}(x)$ по выборке V без предварительной её декомпозиции ($N = 1$) полученные результаты (4), (5) совпадают с асимптотическими выражениями смещения и среднеквадратического отклонения для традиционной непараметрической оценки решающей функции парзеновского типа [5].

Выбор количества составляющих семейства непараметрических решающих функций и оценивание его показателей эффективности. Задача выбора количества N составляющих семейства (2) является первичной при синтезе его структуры и решается на основе результатов аналитических исследований.

Статистика (2) по сравнению с традиционной непараметрической оценкой уравнения разделяющей поверхности

$$\tilde{f}_{12}(x) = \frac{1}{n \prod\limits_{v=1}^{k} c_v} \sum_{i=1}^{n} \prod_{v=1}^{k} \Phi\Big(\frac{x_v - x_v^i}{c_v}\Big) \tag{6}$$

имеет меньшую дисперсию. Поэтому отношение их минимальных асимптотических выражений будем рассматривать в качестве основного критерия при формировании структуры семейства непараметрических оценок решающих функций (2).

Главная составляющая асимптотического выражения дисперсии $\overline{\overline{f}}_{12}(x)$ соответствует первому слагаемому в выражении (5). Его минимальное значение достигается при оптимальных значениях коэффициентов размытости c^* непараметрических оценок составляющих семейства $\overline{\overline{f}}_{12}(x)$.

С учётом результатов работы [4] оптимальные коэффициенты размытости c^* могут быть найдены из условия минимума асимптотического выражения среднеквадратического отклонения

$$\mathrm{M} \int \ldots \int (f_{12}(x) - \bar{f}_{12}^j(x))^2 dx_1 \ldots dx_k \sim \frac{(\bar{n}_1 + \bar{n}_2) \prod\limits_{v=1}^{k} \int \Phi^2(u_v) du_v}{\bar{n}_1 \bar{n}_2 c^k} + \frac{c^4}{4} B.$$

Отсюда нетрудно получить

$$c^* = \left(\frac{k(\bar{n}_1 + \bar{n}_2) \prod\limits_{v=1}^{k} \int \Phi^2(u_v) du_v}{\bar{n}_1 \bar{n}_2 B} \right)^{1/(k+4)}. \tag{7}$$

Тогда, подставляя (7) в первое слагаемое выражения (5), запишем минимальное асимптотическое выражение дисперсии семейства $\overline{\overline{f}}_{12}(x)$ в виде

$$W_3 = \frac{1}{Nk^{k/(k+4)}} \left[\left(\frac{(\bar{n}_1 + \bar{n}_2) \prod\limits_{v=1}^{k} \int \Phi^2(u_v) du_v}{\bar{n}_1 \bar{n}_2} \right)^4 B^k \right]^{1/(k+4)}. \qquad (8)$$

Нетрудно показать, что для традиционной непараметрической решающей функции (8) в принятых условиях $c_v(n) = c$, $v = \overline{1,k}$, оптимальный коэффициент размытости $\tilde{c}(n)$ совпадает со значением $c^*(\bar{n})$ при $\bar{n}_1 = n_1$, $\bar{n}_2 = n_2$. Причём минимальное асимптотическое выражение W_3' дисперсии статистики $\tilde{f}_{12}(x)$ (6) аналогично W_3 при $N = 1$ и $\bar{n}_1 = n_1$, $\bar{n}_2 = n_2$. Здесь n_1 и n_2 — количество элементов обучающей выборки соответственно первого и второго классов ($n_1 + n_2 = n$).

Пусть априорные вероятности классов равны, т. е. $\bar{n}_1 = n_1/N$, $\bar{n}_2 = n_2/N$ и $n_1 = n_2$, тогда отношение

$$R_3 = \frac{W_3}{W_3'} = N^{-k/(k+4)} < 1. \qquad (9)$$

При $k = 1$ отношение R_3 совпадает с результатом работы [1], что подтверждает корректность выполненных преобразований.

Введём пороговое значение $\alpha < 1$ для отношения R_3 и из условия $N^{-k/(k+4)} \le \alpha$ определим количество составляющих семейства (2):

$$\bar{N} = \left| \alpha^{-(k+4)/k} \right| + 1, \qquad (10)$$

а также соответствующее ему значение отношения (9):

$$\bar{R}_3 = \bar{N}^{-k/(k+4)}. \qquad (11)$$

В выражении (10) символом $|\beta|$ обозначена целая часть числа β.

При этом выполняется требуемое условие $\bar{R}_3 < \alpha$ задачи синтеза структуры семейства непараметрических решающих функций, так как с ростом количества его составляющих увеличивается значение $\bar{R}_3$.

Для данного значения $\bar{N}$ вычислим отношение $\bar{R}_2 = W_2/W_2'$ асимптотических выражений среднеквадратических отклонений W_2, W_2' анализируемых оценок решающих функций $\overline{\overline{f}}_{12}(x)$, $\tilde{f}_{12}(x)$ соответственно.

Подставляя оптимальные значения c^* в выражение (5), получим

$$W_2 = \frac{4 + Nk}{4N^{k/(k+4)}} \left[\left(\frac{(\bar{n}_1 + \bar{n}_2) \prod\limits_{v=1}^{k} \int \Phi^2(u_v) du_v}{\bar{n}_1 \bar{n}_2} \right)^4 B^k \right]^{1/(k+4)}. \qquad (12)$$

Для традиционной непараметрической решающей функции (6) минимальное асимптотическое выражение W_2' среднеквадратического отклонения $\tilde{f}_{12}(x)$ от байесовской решающей функции $f_{12}(x)$ может быть получено из W_2 при $N = 1$ и $\bar{n}_1 = n_1$, $\bar{n}_2 = n_2$.

С учётом принятых выше допущений отношение R_2 при $N = \bar{N}$ записывается в виде

$$\bar{R}_2 = \frac{W_2}{W_2'} = \frac{4 + \bar{N}k}{(4+k)\bar{N}^{k/(k+4)}} \tag{13}$$

и при $k = 1$ совпадает с результатом работы [2].

По аналогии вычислим отношение асимптотических выражений смещений W_1, W_1' анализируемых оценок решающих функций $\overline{\overline{f}}_{12}(x)$ и $\tilde{f}_{12}(x)$ при оптимальных коэффициентах размытости ядерных функций и значении $N = \bar{N}$:

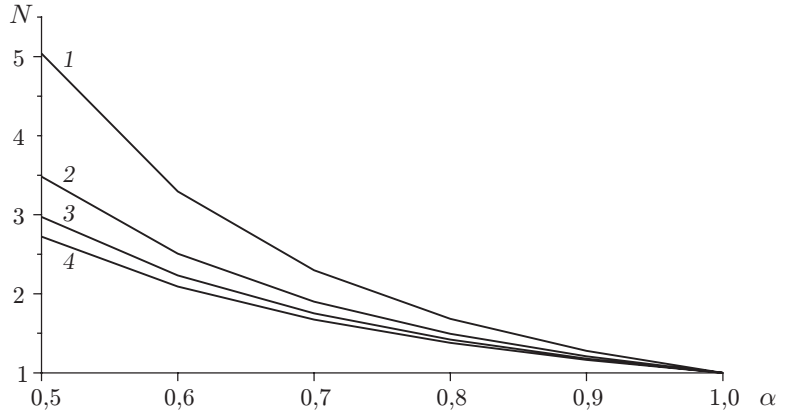
$$\bar{R}_1 = \frac{W_1}{W_1'} = \bar{N}^{2/(k+4)}. \tag{14}$$

Отметим, что отношения $\bar{R}_j$, $j = \overline{1,3}$, получены в результате анализа асимптотических свойств статистик $\overline{\overline{f}}_{12}(x)$ и $\tilde{f}_{12}(x)$. Поэтому применение предлагаемой методики синтеза семейства $\overline{\overline{f}}_{12}(x)$ (2) рекомендуется для исследуемых условий больших выборок, когда использование принципа их декомпозиции является обоснованным.

В качестве примера определим количество $\bar{N}$ составляющих семейства $\overline{\overline{f}}_{12}(x)$ и оценки показателей его эффективности $\bar{R}_j$, $j = \overline{1,3}$, при восстановлении решающей функции в многомерном пространстве признаков (x_v, $v = \overline{1,5}$) классифицируемых объектов. Пороговое значение α для отношения (7) примем равным 0,5.

Для этого обратимся к рисунку и проведём анализ кривой *2*, соответствующей размерности $k = 5$, затем вычислим значение $N(\alpha)$ при $\alpha = 0{,}5$. С учётом рекомендации (10) округлим число $N(\alpha) = 3{,}5$ до целого по верхнему пределу $\bar{N} = 4$. По заданному значению k и количеству $\bar{N}$ составляющих семейства $\overline{\overline{f}}_{12}(x)$ оценим его показатели эффективности, которые определяются значениями отношений $\bar{R}_1 = 1{,}36$, $\bar{R}_2 = 1{,}23$, $\bar{R}_3 = 0{,}46$.

Таким образом, реализация требования уменьшения дисперсии семейства (2) в 2 раза по сравнению с традиционной непараметрической решающей функцией (6) сопровождается увеличением смещения статистики $\overline{\overline{f}}_{12}(x)$ в 1,36 раза и среднеквадратического от-



Зависимости количества N составляющих семейства $\overline{\overline{f}}_{12}(x)$ непараметрических решающих функций от порогового значения α для отношения R_3 и размерности k признаков классифицируемых объектов. Кривые *1–4* соответствуют значениям $k = 3, 5, 7, 9$

клонения в 1,23 раза. При этом семейство $\overline{\overline{f}}_{12}(x)$ обеспечивает сокращение времени формирования значений оценки решающей функции, сопоставимой со значением $\bar{N} = 4$ раза, за счёт возможности использования технологии параллельных вычислений.

Анализ аналитических результатов (11), (13), (14) позволяет выявить особенности аппроксимационных свойств статистики $\overline{\overline{f}}_{12}(x)$.

По сравнению с непараметрической решающей функцией $\tilde{f}_{12}(x)$ предлагаемое семейство $\overline{\overline{f}}_{12}(x)$ обладает меньшей дисперсией, что обусловлено его структурой. Причём с увеличением количества N составляющих семейства $\overline{\overline{f}}_{12}(x)$ и размерности k признаков классифицируемых объектов его преимущество возрастает. При этом наблюдается увеличение смещения и среднеквадратического отклонения $\overline{\overline{f}}_{12}(x)$, что объясняется снижением объёмов выборок при оценивании решающих функций $\bar{f}_{12}^{j}(x)$, $j = \overline{1, N}$. Данная тенденция особенно характерна для малых размерностей k случайной величины x.

Заключение. Анализ асимптотических свойств семейства непараметрических решающих функций, основанных на оценках плотности вероятности многомерных случайных величин типа Розенблатта — Парзена, проведённый в представленной работе, позволяет обосновать подход к синтезу и анализу его структуры в условиях больших выборок. Предложена методика выбора количества составляющих семейства и оценивания показателей его эффективности. Исследуемая статистика по сравнению с традиционной непараметрической решающей функцией имеет меньшую дисперсию и позволяет использовать технологию параллельных вычислений.

СПИСОК ЛИТЕРАТУРЫ

1. **Лапко А. В., Лапко В. А.** Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. 2010. **46**, № 1. С. 70–78.
2. **Лапко А. В., Лапко В. А.** Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. 2009. **37**, № 3.1. С. 156–160.
3. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
4. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, № 1. С. 156–161.
5. **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.

Поступила в редакцию 31 января 2011 г.