УДК 621.391:519.28

СИНТЕЗ НЕПАРАМЕТРИЧЕСКИХ АЛГОРИТМОВ РАСПОЗНАВАНИЯ ОБРАЗОВ, ЗАДАННЫХ СЛОЖНЫМИ ЭТАЛОННЫМИ ОПИСАНИЯМИ, НА ОСНОВЕ АППРОКСИМАЦИИ СТАТИСТИЧЕСКИХ УСРЕДНЕННЫХ ФУНКЦИЙ ПРАВДОПОДОБИЯ ПОЛИГОНАМИ СМИРНОВА

д.т.н. Г.В. Певцов, В.Л. Русских, В.А. Лупандин

Разработан метод синтеза непараметрических алгоритмов распознавания образов, заданных сложными эталонными описаниями в виде совокупностей эталонных интервалов и (или) дискретных значений признаков. Обучение алгоритма основано на получении полигональных оценок статистических усредненных функций правдоподобия по Смирнову.

Постановка проблемы и анализ литературы. При разработке аппаратуры диагностики и контроля физических, химических, биологических и других процессов часто возникает необходимость синтеза алгоритмов распознавания образов, заданных сложными эталонными описаниями в виде интервалов возможных значений и (или) совокупностей дискретных эталонных значений признаков (параметров процессов). Для решения задач такого типа в [1 – 4] и др. разработана методика синтеза алгоритмов распознавания на основе проверки сложных статистических гипотез путем сравнения статистик отношений усредненных функций правдоподобия с порогом. При синтезе алгоритмов, реализующих разработанную методику, эталонные описания образов определяются в предположении о том, что виды и параметры априорных распределений известны. Однако в практике создания систем распознавания образов чаще возникает необходимость в алгоритмах, которые могут обучаться до или в процессе ведения распознавания. Одним из наиболее простых, но достаточно эффективных методов обучения является построение сравниваемых статистик гистограммным методом с последующим выравниванием полигонами Смирнова [5].

Целью статьи является развитие методики синтеза алгоритмов распознавания образов, заданных сложными эталонными описаниями, на случай непараметрического обучения путем построения по Смирнову полигональных оценок усредненных статистических функций правдоподобия.

Пусть на множестве $\mathbf{U}$ объектов распознавания наблюдается L образов $\mathbf{U}_i \subset \mathbf{U}$, $i \in \{1, 2, \dots, L\}$, представляющих собой множества (группы) объектов распознавания. Каждый из образов проявляется в метрике $\Im$-мерного евклидового пространства признаков $\mathbf{S}$. В метрике каждого признака s_j, $j \in \{1, 2, \dots, \Im\}$, каждому образу априорно может соответствовать ν_i элементов — эталонных значений и (или) интервалов эталонных значений признака, описывающих группу объектов распознавания. Неизвестное априорное распределение вектора признаков $\mathbf{s} = \{s_1, s_2, \dots, s_j, \dots, s_\Im\}$ для каждого i-го образа представляет собой $\Im$-мерную совместную плотность вероятности смешанного типа $W(\mathbf{s} \mid \mathbf{U}_i) = w_i(\mathbf{s})$ вектора $\mathbf{s}$ на множестве $\mathbf{U}_i$, определенную в области $\mathbf{S}_i$ пространства признаков. По результатам Ξ испытаний для каждого из образов в метрике каждого признака s_j, $j \in \{1, 2, \dots, \Im\}$, получены первичные статистические совокупности, из которых сформированы группированные статистические ряды. Выдвигается L гипотез $H_1, H_2, \dots, H_L$ о том, что наблюдаемая выборка $\mathbf{x}$ ζ-кратно измеренных значений $\Im$ признаков принадлежит одному из образов $\mathbf{U}_i$. Задача состоит в установлении до наблюдения дискретно-аналогового нерандомизированного статистически оптимального правила δ, реализующего разделение $\zeta \times \Im$-мерного евклидового пространства выборок $\mathbf{X}$ на L непересекающихся областей $\mathbf{X}^*_q$, $q \in \{1, 2, \dots, L\}$, $\bigcup_{q=1}^{L} X^*_q = X$, и приписывающего каждой из этих областей одного из L решений γ_q о принятии гипотезы H_q.

Опишем аналитически полигональную оценку w^*_{ijn} произвольной одномерной эмпирической дифференциальной функции распределения, соответствующей n-му элементу, $n \in \{1, 2, \dots, \nu_i\}$, i-го образа в метрике j-го признака, в виде [5]:

$$w^*_{ijn}(x_j) = \frac{1}{\Xi_{ijn} \Delta s_j} \left| 0{,}5\left(\Xi_{ijnk} + \Xi_{ijn(k+1)}\right) + \frac{1}{\Delta s_j}\left(x_j - k_j \Delta s_j\right)\left(\Xi_{ijn(k+1)} - \Xi_{ijnk}\right) \right|; \quad (1)$$

$$k_j = \mathrm{ent}\,(x_j/\Delta s_j); \qquad x_j \in [k_j \Delta s_j - 0{,}5\Delta s_j,\ k_j \Delta s_j + 0{,}5\Delta s_j],$$

где Ξ_{ijnk} – количество выборочных значений j-го признака, попавших в k-й интервал $[(k_j-1)\Delta s_j,\ k_j \Delta s_j]$ при наблюдении (в ходе испытаний) n-го элемента i-го образа; Ξ_{ijn} – количество наблюдений n-го элемента i-го образа в метрике j-го признака; $\mathrm{ent}(\cdot)$ – символ операции округления.

Статистики i-х элементов

$$\Lambda_i(x) = w_i(x)/w_1(x) \qquad (2)$$

вектора отношений усредненных статистических функций правдоподобия [5]:

$$w_i(x) = W(x|U_i) = \int_{S_i} w_i(s)W(x \mid s)ds \qquad (3)$$

определим в соответствии с [1 – 4], полагая признаки и наблюдаемые выборки независимыми:

$$w_i(x) = \prod_{j=1}^{\Im}\left[\sum_{r=1}^{R_{ij}} I_{ijr}p_{ijr}\int_{S_{ij}} w_{ijr}\left(s_j, s'_{ijr}, s''_{ijr}\right)W\left(x \mid s_j\right)ds_j + \sum_{d=1}^{D_{ij}} I_{ijd}p_{ijd}W\left(x \mid s_{ijd}\right)\right]; \qquad (4)$$

$$\sum_{r=1}^{R_{ij}} p_{ijr} + \sum_{d=1}^{D_{ij}} p_{ijd} = 1, \ \sum_{r=1}^{R_{ij}} I_{ijr}p_{ijr} + \sum_{d=1}^{D_{ij}} I_{ijd}p_{ijd} = 1,$$

$$\forall i \in \{1, 2, \dots, L\}, \quad \forall j \in \{1, 2, \dots, \Im\},$$

где $w_{ijr}(s_j, s'_{ijr}, s''_{ijr})$ – априорные плотности распределения признака s_j на каждом из R_{ij} эталонных интервалов $[s'_{ijr}, s''_{ijr}]$, $r\in\{1, 2, \dots , R_{ij}\}$; $\delta(s_j - s_{ijd})$ – функции Дирака, как плотности вероятности математических ожиданий s_{ijd} каждого из D_{ij} возможных дискретных эталонных значений признака s_j, $d\in\{1, 2, \dots , D_{ij}\}$; p_{ijr} и p_{ijd} – априорные условные вероятности наблюдения r-го интервала или d-го значения при наблюдении образа $\mathbf{U}_i$ в метрике признака s_j; $I_{ijr(d)}\in[0, 1]$ – коэффициенты, характеризующие относительную степень информативности r-го интервала или d-го значения признака s_j при распознавании образа $\mathbf{U}_i$; $W(\mathbf{x} \mid \mathbf{s})$ — зависящая от значений вектора параметров $\mathbf{s}$ функция правдоподобия наблюдаемой выборки $\mathbf{x}$; $\mathbf{S}_i$ – область определения образа $\mathbf{U}_i$ в пространстве признаков $\mathbf{S}$, $\mathbf{S}_{ij}$ – область определения образа $\mathbf{U}_i$ в метрике признака s_j.

Для нахождения оценок (2), (3) подставим в (4) вместо

$\int_{S_{ij}} w_{ijr}(s_j, s'_{ijr}, s''_{ijr})W(x \mid s_j)\, ds_j$, $W(x \mid s_{ijd})$, p_{ijr} и p_{ijd} их оценки.

Первые два элемента получим из (1) заменой индексов n на r и d, соответственно. Оценки априорных вероятностей p_{ijr} и p_{ijd} определим отношениями:

$$p^*_{ijr} = \Xi_{ijr}/\Xi_i\,;$$

$$p^*_{ijd} = \Xi_{ijd}/\Xi_i$$

количества проявлений r-го интервала Ξ_{ijr} и d-го значения Ξ_{ijd} i-го образа в метрике j-го признака к количеству Ξ_i наблюдений i-го образа. В результате из (1), (4) имеем:

$$w^*_i(x)=\frac{1}{\Xi_i^{\Im}}\prod_{j=1}^{\Im}\frac{1}{\Delta s_j}\left\{\sum_{r=1}^{R_{ij}}\frac{I_{ijr}}{\varpi_{ijr}}\Big[0{,}5\big(\Xi_{ijrk}+\Xi_{ijr(k+1)}\big)+\frac{1}{\Delta s_j}\big(x_j-k_j\Delta s_j\big)\big(\Xi_{ijr(k-1)}-\right.$$

$$\left.-\Xi_{ijrk}\big)\Big]+\sum_{d=1}^{D_{ij}}\frac{I_{ijd}}{\varpi_{ijd}}\left[0{,}5\big(\Xi_{ijdk}+\Xi_{ijd(k+1)}\big)+\frac{1}{\Delta s_j}\big(x_j-k_j\Delta s_j\big)\big(\Xi_{ijd(k+1)}-\Xi_{ijdk}\big)\right]\right\};\quad(5)$$

$$k_j=\mathrm{ent}(x_j/\Delta s_j);\qquad x_j\in\big[k_j\Delta s_j-0{,}5\Delta s_j,\ k_j\Delta s_j+0{,}5\Delta s_j\big],$$

где $\varpi_{ijn}=\int\limits_{-\infty}^{\infty}w^*_{ijn}(x_j)\,dx_j$ – коэффициент, с помощью которого компенсируется возможное (из-за погрешностей интерполяции) отличие площади под ломаной (1) от единичного значения.

Подставляя (5) в (2), получаем полигональные оценки отношений правдоподобия для непараметрических алгоритмов распознавания образов, заданных сложными эталонными описаниями. Синтезируемые алгоритмы распознавания предполагают сравнения полученных отношений правдоподобия с порогом, значение которого определяется выбранным критерием эффективности [1 – 4] с учетом оценок p^*_i априорных вероятностей наблюдения образов $\mathbf{U}_i$, $p^*_i=\Xi_i/\Xi$, $\sum\limits_{i=1}^{L}p^*_i=1$.

При анализе синтезированных алгоритмов необходимо получить оценку полной вероятности ошибки $p_{ош}$ вида

$$p^*_{ош}=\sum_{i=1}^{L}p^*_{ош\,i}=\sum_{i=1}^{L}p^*_i\sum_{\substack{q=1\\q\neq i}}^{L}P^*\{\gamma_q|U_i\};\quad \bigcup_{i=1}^{L}X^*_i=X;\quad \bigcap_{i=1}^{L}X^*_i=\varnothing.\qquad(6)$$

Входящие в (6) оценки $P^*\{\gamma_q|\mathbf{U}_i\}$ полных вероятностей принятия ошибочных решений γ_q при наблюдении образа $\mathbf{U}_i$ определяются как усредненные по эталонным описаниям образов оценки условных вероятностей ошибок [3, 4]:

$$P^*\{\gamma_q|\mathbf{U}_i\}=\int\limits_{X^*_q}w^*_i(x)\,dx\,.\qquad(7)$$

Из (5) – (7) имеем асимптотическую оценку полной вероятности ошибки:

$$p^*_{ош}=\frac{1}{\Xi}\sum_{i=1}^{L}\frac{1}{\Xi_i^{\Im-1}}\sum_{\substack{q=1\\q\neq i}}^{L}\int\limits_{X^*_q}\prod_{j=1}^{\Im}\frac{1}{\Delta s_j}\left\{\sum_{r=1}^{R_{ij}}\frac{I_{ijr}}{\varpi_{ijr}}\left[0{,}5\big(\Xi_{ijrk}+\Xi_{ijr(k+1)}\big)+\frac{1}{\Delta s_j}\big(x_j-k_j\Delta s_j\big)\times\right.\right.$$

$$\times\left(\Xi_{ijr(k+1)}-\Xi_{ijrk}\right)\Big]+\sum_{d=1}^{D_{ij}}\frac{I_{ijd}}{\varpi_{ijd}}\left[0{,}5\left(\Xi_{ijdk}+\Xi_{ijd(k+1)}\right)+\frac{1}{\Delta s_j}\left(x_j-k_j\Delta s_j\right)\left(\Xi_{ijd(k+1)}-\Xi_{ijdk}\right)\right]\Bigg]\Bigg\}dx\,; \quad (8)$$

$$k_j=\mathrm{ent}(x_j/\Delta s_j);\qquad x_j\in\left[k_j\Delta s_j-0{,}5\Delta s_j,\ k_j\Delta s_j+0{,}5\Delta s_j\right].$$

Выводы. Таким образом, разработанный метод позволяет синтезировать непараметрические алгоритмы распознавания образов, заданных сложными эталонными описаниями, с обучением путем получения первичных статистических совокупностей и формирования из них отношений усредненных статистических функций правдоподобия, выровненных полигонами Смирнова.

ЛИТЕРАТУРА

1. *Певцов Г.В. Синтез алгоритма распознавания радиоизлучений на основе байесовского правила проверки сложных гипотез // Радиоэлектроника (Изв. высш. учебн. заведений). – 1998. – № 4. – С. 49 – 57.*
2. *Певцов Г.В. Синтез алгоритмов распознавания образов, заданных сложными эталонными описаниями в метрике азимутов на источники радиоизлучений // Радиоэлектроника (Изв. высш. учебн. заведений). – 2000. – № 4. – С. 38 – 45.*
3. *Певцов Г.В., Лупандин В.А. Синтез алгоритмов многоальтернативного распознавания образов на основе проверки сложных статистических гипотез по критерию максимума апостериорной вероятности // Радиоэлектроника (Изв. высш. учебн. заведений).– 2001. – № 11. – С. 77 – 80.*
4. *Певцов Г.В. Синтез байесовских алгоритмов многоальтернативного распознавания образов, заданных сложными эталонными описаниями // Радиоэлектроника (Изв. высш. учебн. заведений). – 2003. – № 1. – С. 58 – 63.*
5. *Фомин Я.А., Тарловский Г.Р. Статистическая теория распознавания образов. – М.: Радио и связь, 1986. – 264 с.*
6. *Левин Б.Р. Теоретические основы статистической радиотехники. Книга 2. – М.: Сов. радио, 1975. – 392 с.*

Поступила 26.06.2003

***ПЕВЦОВ Геннадий Владимирович**, доктор техн. наук, ст. научн. сотр., зам. начальника научного центра при ХВУ по научной работе. В 1983 году окончил Киевское ВИРТУ ПВО. Область научных интересов – обработка радиотехнических сигналов и информации.*

***РУССКИХ Виктор Леонидович**, начальник отдела – заместитель начальника Главного штаба Войск ПВО по военно-научной работе. В 1989 году окончил ВА ПВО СВ им. А.М. Василевского. Область научных интересов – военная кибернетика.*

***ЛУПАНДИН Владимир Анатольевич**, адъюнкт ХВУ. В 1992 году окончил Харьковское ВВКИУРВ. Область научных интересов – обработка радиотехнических сигналов и информации.*