УДК 621.3911:519.28

И. К. ВАСИЛЬЕВА

Национальный аэрокосмический университет им. Н. Е. Жуковского "ХАИ", Украина

ИТЕРАЦИОННЫЙ МЕТОД ОЦЕНКИ ПАРАМЕТРОВ СМЕСИ ФУНКЦИЙ ГАУССА В ЗАДАЧАХ ОПИСАНИЯ ДАННЫХ НАБЛЮДЕНИЙ

Предложена математическая модель в виде смеси базовых функций для описания форм кривых, заданных массивами выборок. Описана методика оценки параметров модели, которая состоит в последовательном уточнении множеств векторов оценок параметров, полученных на основе информации об ошибках аппроксимации на каждом шаге итерационной процедуры, и даны основные расчетные соотношения. Приведены результаты исследования точности описания контрольных массивов с различной структурой взаимосвязи данных смесями функций Гаусса с использованием предлагаемого метода оценки параметров модели. Показано, что данный метод является эффективным и может применяться для определения количества компонент смеси функций Гаусса и оценок параметров этих компонент при описании массивов данных в задачах распознавания образов и кластерного анализа.

Ключевые слова: *распознавание образов, описание формы, граница образа, параметрически заданная кривая, аппроксимация, смесь базовых функций, оценка параметров смеси, итерация, критерий точности.*

Введение

Реализация методов распознавания необходима в автоматизированных системах, использующих возможности искусственного интеллекта, предназначенных для решения задач диагностики, мониторинга, прогнозирования, обучения, управления поведением сложных систем. Такие методы теории распознавания, как кластерный анализ, выявление закономерностей в экспериментальных данных, прогнозирование различных процессов или явлений широко используются в научных исследованиях. Выбор признаков для распознавания связан с выделением атрибутов, которые выражают интересующую количественную информацию и являются основой классификации объектов. В тех случаях, когда интерес представляют характеристики формы объектов, для представления зрительных образов используют описание границ областей, соответствующих изображениям объектов [1, 2]. Такой подход позволяет исключить из рассмотрения внутренние точки изображения и тем самым значительно сократить объем хранимой и обрабатываемой информации, что часто позволяет обеспечить работу системы принятия решений в режиме реального времени. Вероятность распознавания пространственных объектов по их изображениям в значительной мере зависит от сохранения подобия контуров изображения и оригинала с учетом различных шумов и искажений, возникающих при формировании и обработке цифровых изображений.

Наиболее общими подходами к представлению границ объекта являются аппроксимация кривых, прослеживание контуров и связывание точек перепадов яркости. Основная задача при этом состоит в формировании по двумерной форме объекта описания его границы с помощью некоторой одномерной функции $g(x)$. Примерами таких функций являются функция тангенциального представления угла, противоположного отрезку дуги границы (функция поворота), комплексная функция $\dot{x}(t) = jy(t)$, где t – длина дуги. В сигнатурном анализе используется представление функции контура относительно центра тяжести, при этом существуют следующие варианты: функция точек границы с равномерным шагом выборки $\Delta t = \text{const}$, функция углового поворота при равном расстоянии между точками, замена границы n-полигоном с равными сторонами с последующим вычислением функции отношения угла между стороной и радиус-вектором к этой стороне, функция расстояния от центра до точек с высокой кривизной [3]. Также для описания границ применяют ряды Фурье, регрессионные и графовые модели, цепные коды и др. В структурном анализе растровых изображений границы обычно описывают в виде последовательности особых точек или отрезков прямых [4]. Известно также, что в большом количестве приложений криволинейные элементы изображений представляют в виде сплайнов, кривых Безье и т. п.

В данной работе для описания координат границы (x_i, y_i) использован метод аппроксимации одномерных массивов $\{x_i\}$ и $\{y_i\}$ смесью функций Гаусса. Смесь базовых функций является одной из распространенных математических моделей, используемых в задачах классификации без обучения [5, 6], когда каждый класс интерпретируется как параметрически заданная одномодальная генераль-

ная совокупность (при неизвестном значении определяющего ее параметра), а классифицируемые наблюдения – как выборка из смеси таких совокупностей; при этом число компонент смеси трактуется как количество классов, а удельных веса этих компонент – как их априорные вероятности. Основным методом решения задачи расщепления смеси является статистический анализ, проводимый в рамках одной из двух логических схем. В первой из них реализуется логика «от оценивания параметров смеси к классификации» (EM-алгоритмы [7, 8], основанные на методе максимального правдоподобия, методе моментов и т. д.). Во второй, напротив, идут «от классификации к оцениванию»: выбрав начальное разбиение выборочного множества на классы и получив оценки параметров распределений внутри классов, уточняют классификацию и т. д. (алгоритм SEM адаптивного вероятностного обучения [7, 8]).

В данной работе предложен итерационный метод оценки параметров смеси по массиву неотрицательных данных $\{f_i\}^{(r)}$, r = 0 (номер итерации), основанный на последовательном нахождении значений элементов $\{f_i\}$ в малых окрестностях локальных максимумов, численном вычислении по этим значениям первых производных базисных функций, расчете по полученным данным оценок параметров смеси, определении массива ошибок аппроксимации $\{d_i\}$ и возвращению к началу процедуры с заменой $r = r+1$, $\{f_i\}^{(r)} = \{d_i\}$ до тех пор, пока не будет удовлетворено условие остановки счета.

Целью работы является исследование применимости предлагаемого метода для описания выборочных данных одномерной функцией в задачах классификации образов и группировки данных.

1. Описание метода

Одной из распространенных моделей для описания неизвестных функциональных зависимостей является представление их в виде смести базовых функций. В данной работе в качестве базовых были приняты ненормированные функции Гаусса

$$N(z) = \exp\left[-\frac{1}{2}\left(\frac{z-m}{\sigma}\right)^2\right], \qquad (1)$$

где m – параметр сдвига по координате z, определяющий положение моды функции, $N(m) = 1$;

σ – параметр масштаба, характеризующий скорость убывания функции от ее модального значения в точке $z = m$, $N(m \pm 3\sigma) = 0{,}011$.

Рассматриваемая модель имеет вид

$$s\left(z\middle|\vec{A}, \vec{m}, \vec{\sigma}\right) = \sum_{k=1}^{K} A_k N(z; m_k, \sigma_k), \qquad (2)$$

где K – количество компонент смеси;

A_k – весовой коэффициент k-й компоненты;

m_k и σ_k – параметры сдвига и масштаба k-й базовой функции, соответственно.

Т. о., оценке подлежат следующие параметры смеси: $\left\{K, \vec{A}, \vec{m}, \vec{\sigma}\right\}$. В основу метода оценивания значений параметров m и σ k-й компоненты смеси (нижние индексы условно опущены) положено следующее тождество:

$$\frac{dN(z)}{dz} = -\frac{z-m}{\sigma^2} N(z). \qquad (3)$$

Для нахождения двух неизвестных требуется решить двухточечную задачу, т. е. записать уравнение (3) для двух соседних отсчетов обучающей выборки. При этом необходимо заменить в (3) производную на конечно-разностное выражение

$$\frac{dN(z)}{dz} = \frac{N(z+\Delta z) - N(z-\Delta z)}{2\Delta z}. \qquad (4)$$

При аппроксимации массива данных $\{f_i\}$, $i = 0, \ldots, M-1$ такая система имеет вид

$$\begin{cases} \frac{1}{2}(f_{i+1} - f_{i-1}) = \frac{m-i}{\sigma^2} \cdot f_i; \\ \frac{1}{2}(f_{i+2} - f_i) = \frac{m-i-1}{\sigma^2} \cdot f_{i+1}, \end{cases} \qquad (5)$$

для $i = 1, \ldots, M-2$.

Корни системы (5):

$$\hat{m}_i = i + \frac{f_{i+1}^2 - f_{i+1} \cdot f_{i-1}}{-f_{i+2} \cdot f_i + f_{i+1}^2 - f_{i+1} \cdot f_{i-1} + f_i^2}; \qquad (6)$$

$$\hat{\sigma}_i = \sqrt{\left|\frac{2f_i \cdot (m_i - i)}{f_{i+1} - f_{i-1}}\right|} \qquad (7)$$

являются множеством оценок параметров m и σ.

Точность оценок m_k, σ_k зависит от погрешностей, вносимых как конечно-разностным представлением производной, так и влиянием остальных (j = 1…K, j ≠ k) компонент смести; лучшие результаты были получены при индексах отсчетов, близких к оцениваемым значениям параметра m (т. е. в точках локальных максимумов выборочных данных). Поэтому расчеты по (6), (7) следует проводить только для подмножества индексов $j \subset i$, соответствующих положениям локальных максимумов, а также граничным точкам массива $\{f_i\}$, если те являются верхними гранями подмножеств, образованных своими малыми окрестностями:

$$j = \begin{cases} 1, & \text{если } f_0 > f_1; \\ i, & \text{если } f_{i-1} < f_i > f_{i+1}, i = 1, \ldots, M-2; \\ M-2, & \text{если } f_{M-1} > f_{M-2}. \end{cases} \qquad (8)$$

Мощность подмножества индексов $\{j\}$ является оценкой количества компонент смеси, выявленных на текущем этапе.

После определения состава подмножества $\{j\}$ по формулам (6), (7) вычисляются оценки m_j и σ_j. Если рассчитанное значение σ_ν, $\nu \in \{j\}$ оказывается меньше порогового значения ε_σ (в данной работе $\varepsilon_\sigma = 10^{-4}$), то его переопределяют как

$$\sigma_\nu = \begin{cases} \sigma_{\nu+1}, & \text{если } \nu = 1; \\ \dfrac{\sigma_{\nu+1} + \sigma_{\nu-1}}{2}, & \text{если } 1 < \nu < M-2; \\ \sigma_{\nu-1}, & \text{если } \nu = M-2. \end{cases} \quad (9)$$

Поскольку $A_j \cdot N(m_j; m_j, \sigma_j) = A_j$ и положения локальных максимумов m_j (для внутренних элементов массива $\{f_i\}$) определяются значениями индексов из подмножества $\{j\}$:

$$j = \text{round}(m_j),$$

где round(•) – функция округления вещественного числа, то предварительными оценками весовых коэффициентов служат элементы массива $\{f_i\}$ с индексами $\{j\}$:

$$A_j = f_j. \quad (10)$$

Если значения m_j оцениваются по граничным точкам, то эти оценки могут выходить за пределы допустимого множества индексов элементов массива $\{f_i\}$ (т. е. $\text{round}(m_j) \notin [0, M-1]$). В этом случае применима формула

$$A_j = \begin{cases} \dfrac{f_0}{N(0; m_j, \sigma_j)}, \text{если round}(m_j) < 0; \\ \dfrac{f_{M-1}}{N(M-1; m_j, \sigma_j)}, \text{если round}(m_j) > M-1. \end{cases} \quad (11)$$

Полученные по (10) или (11) оценки коэффициентов A_j уточняют таким образом:

$$A_j = A_j - \sum_{k, k \neq j} A_k \cdot N(m_j; m_k, \sigma_k). \quad (12)$$

Перечисленные выше этапы позволяют определить начальные приближения параметров смеси, которые обозначим K_0, $\vec{m}^{(0)}$, $\vec{\sigma}^{(0)}$, $\vec{A}^{(0)}$.

Получив массив значений модели (2)

$$s_i^{(0)} = \sum_{k=1}^{K_0} A_k^{(0)} N\left(i; m_k^{(0)}, \sigma_k^{(0)}\right), \quad (13)$$

можно вычислить значение критерия точности модели, например, сумму квадратов отклонений

$$E_0 = \sum_i \left(f_i - s_i^{(0)}\right)^2. \quad (14)$$

Если точность описания данных моделью (2) при начальных приближениях параметров смеси является неудовлетворительной (при принятом критерии качества аппроксимации), то формируют разностный массив

$$d_i = f_i - s_i^{(0)}; \quad (15)$$

такие значения массива $\{d_i\}$, которые по абсолютной величине меньше порога ε_d считаются незначимыми и обнуляются.

Данный алгоритм оценки параметров смеси ориентирован на поиск локальных максимумов и преобразование точек их ближайшей окрестности (если рассматривать массив данных $\{f_i\}$ как таблично заданную неизвестную функцию), поэтому, если в разностном массиве $\{d_i\}$ все значимые по величине элементы неотрицательны, то $\{d_i\}$ является дополнением к модели (2), полученной на текущем этапе; т. о., массив $\{d_i\}$ заменяет первоначальный массив $\{f_i\}$ и итерационная процедура оценки параметров смеси повторяется. В противном случае необходимо переопределить некоторые из оценок параметров смеси. Например, можно пересчитать оценки коэффициентов $A_j^{(0)}$:

$$A_j^{(0)} = A_j^{(0)} + d_{\min}, \quad (16)$$

где $d_{\min}$ – величина минимальный элемент разностного массива, $d_{\min} < 0$.

В ряде случаев более эффективным является перечет оценок параметра σ по правилу:

$$\sigma_k = \begin{cases} \sigma_k, \text{если } \sigma_k < \dfrac{|\mu - m_k|}{\sqrt{2 \ln\left(\left|\dfrac{0{,}01}{A_k}\right|\right)}}; \\ \dfrac{|\mu - m_k|}{\sqrt{2 \ln\left(\left|\dfrac{0{,}01}{A_k}\right|\right)}}, \text{в противном случае,} \end{cases} \quad (17)$$

где μ – индекс элемента $d_{\min}$.

После обеспечения требования $\forall i : d_i \geq 0$ процедура оценки параметров смеси, описанная выше, применяется к разностному массиву $\{d_i\}$, что позволяет выявить скрытые компоненты смеси и уточнить оценки параметров смеси по формулам (6) – (12). Результаты, полученные на этом шаге, обозначим K_r, $\vec{m}^{(r)}$, $\vec{\sigma}^{(r)}$, $\vec{A}^{(r)}$ при $r = 1$, где r – номер шага итерационной процедуры.

Уточнение оценки количества компонент смеси состоит в поэлементном сравнении векторов $\vec{m}^{(r-1)}$ и $\vec{m}^{(r)}$. Если величины отдельных компонент этих векторов отличаются незначительно (например, не более чем на 5 %), то они усредняются и суммарное количество компонент смеси уменьшается на число усредненных оценок $\vec{m}$. Т. о., если для всех индексов $\nu \in \{j\}^{(r-1)}$, $\eta \in \{j\}^{(r)}$ выполняется условие:

$$\frac{\left|m_\nu^{(r-1)} - m_\eta^{(r)}\right|}{\min\left\{m_\nu^{(r-1)}, m_\eta^{(r)}\right\}} < \delta_m, \quad (18)$$

где δ_m – принятое предельно допустимое значение погрешности оценки параметра m, то

$$m_\eta^{(r)} = \frac{1}{2}\left(m_\nu^{(r-1)} + m_\eta^{(r)}\right), \qquad (19)$$

$$\sigma_\eta^{(r)} = \sqrt{\frac{A_\eta^{(r)}}{A_\eta^{(r)} + A_\nu^{(r-1)}}\sigma_\eta^{(r)^2} + \frac{A_\nu^{(r-1)}}{A_\eta^{(r)} + A_\nu^{(r-1)}}\sigma_\nu^{(r-1)^2}}, \qquad (20)$$

$$A_\eta^{(r)} = A_\eta^{(r)} + A_\nu^{(r-1)}. \qquad (21)$$

Если условие (18) не выполняется, то вектор оценок $\vec{m}^{(r)}$ увеличивается на компоненту $m_\nu^{(r-1)}$. При этом в $\vec{\sigma}^{(r)}$ включают $\sigma_\nu^{(r-1)}$, а в $\vec{A}^{(r)}$ – $A_\nu^{(r-1)}$.

После уточнения мощности множеств оценок параметров компонент смеси (на r-м шаге) и значений их элементов формируется массив $\left\{s_i^{(r)}\right\}$

$$s_i^{(r)} = \sum_{k=1}^{K_r} A_k^{(r)} N\left(i; m_k^{(r)}, \sigma_k^{(r)}\right)$$

и вычисляется новое значение критерия точности модели E_r. Если $E_r > \varepsilon_a$, где ε_a – допустимая погрешности аппроксимации и при этом $E_r < E_{r-1}$, то определяют массивы значений

$$d_i = f_i - s_i^{(r)},$$

$$A_j^{(r)} = A_j^{(r)} + \min\{d\},$$

и пересчитывают $\left\{s_i^{(r)}\right\}$ и $\{d_i\}$.

Разностный массив $\{d_i\}$ является объектом аппроксимации на следующем шаге итерационной процедуры, $r = r + 1$. Если приемлемая точность описания данных моделью (2) достигнута, то процедуру прекращают. Альтернативным критерием остановки счета может служить превышение заданного количества итераций.

2. Результаты проверки эффективности метода оценки параметров модели

В качестве тестовых данных на первом этапе апробации предлагаемого метода описания данных были взяты массивы значений, рассчитанных по модели (2), т. о., модель (2) априори полностью соответствовала структуре взаимосвязи данных в контрольных выборках, а путем сравнения оценок параметров смеси K, $\vec{m}$, $\vec{\sigma}$, $\vec{A}$ с их действительными значениями можно было установить эффективность процедуры в смысле обеспечения требуемой точности оценок. Значения параметров компонент смеси (2) для формирования контрольных массивов $\{f1_i\}$, $\{f2_i\}$, $\{f3_i\}$, $i = 0,\ldots,M-1$ (M = 120) приведены в табл. 1; там же указаны величины критерия точности модели E_2 (14), достигнутые на втором шаге итерационной процедуры.

На рис. 1 показаны графики соответствующих контрольных массивов и массивов значений модели (2) с оценками параметров смеси $\vec{m}_1^{(2)}$, $\vec{\sigma}_1^{(2)}$, $\vec{A}_1^{(2)}$.

Количество компонент смеси во всех приведенных случаях K = 4; начальное приближение для K определялось по количеству локальных максимумов; т. о., значение K_0 составило 4 (для $\{f1_i\}$), 2 (для $\{f2_i\}$) и 1 (для $\{f3_i\}$). При этом уже на следующем шаге оценка K (размерность векторов оценок $\vec{m}$, $\vec{\sigma}$, $\vec{A}$) для всех контрольных массивов стала равной действительному значению компонент смеси K = 4.

Как видно из результатов, представленных в табл. 1, для массива $f1_i = s\left(i\middle|\vec{A}_1, \vec{m}_1, \vec{\sigma}_1\right)$ (рис. 1, а) уже за два шага достигается приемлемая точность как описания формы кривой, так и определения значений параметров смеси; так, максимальные относительные погрешности оценок соответствующих параметров модели (2) составляют: $\max\{\delta_m\} < 1$ %, $\max\{\delta_\sigma\} < 15$ %, $\max\{\delta_A\} < 1$ %. Для других случаев (когда количество локальных максимумов меньше числа компонент смеси, см. рис. 1, б, в) при удовлетворительном совпадении графиков контрольных массивов и графиков результатов их аппроксимации моделью (2), что подтверждается относительно малыми значениями критерия E_2 (см. табл. 1), оценки параметров смеси $\vec{m}_1^{(2)}$, $\vec{\sigma}_1^{(2)}$, $\vec{A}_1^{(2)}$ существенно отличались от их действительных значений; так, относительные погрешности оценок составили δ_m = (1…75) %, δ_σ = (15…150) %, δ_A = (20…125) %.

Таблица 1

Значения параметров компонент тестовых массивов и их оценки (количество итераций r = 2)

$\vec{m}_1$	$\vec{m}_1^{(2)}$	$\vec{\sigma}_1$	$\vec{\sigma}_1^{(2)}$	$\vec{A}_1$	$\vec{A}_1^{(2)}$	E_2
5	5,032	5	5,656	2	2,01	0,211
25	25,118	8	7,954	3	2,987	
50	50,127	7	7,162	3	3,001	
75	75,083	9	8,966	1	0,997	
$\vec{m}_2$	$\vec{m}_2^{(2)}$	$\vec{\sigma}_2$	$\vec{\sigma}_2^{(2)}$	$\vec{A}_2$	$\vec{A}_2^{(2)}$	E_2
5	8,741	6	12,513	1	2,185	0,659
18	24,395	12	7,444	2	0,679	
50	50,809	10	11,532	3	3,525	
70	79,406	20	12,914	1	0,787	
$\vec{m}_3$	$\vec{m}_3^{(2)}$	$\vec{\sigma}_3$	$\vec{\sigma}_3^{(2)}$	$\vec{A}_3$	$\vec{A}_3^{(2)}$	E_2
10	3,878	12	6,404	3	0,36	0,458
25	25,96	8	19,605	2	4,414	
40	44,487	10	7,054	3	0,482	
70	77,549	20	14,87	1	0,795	

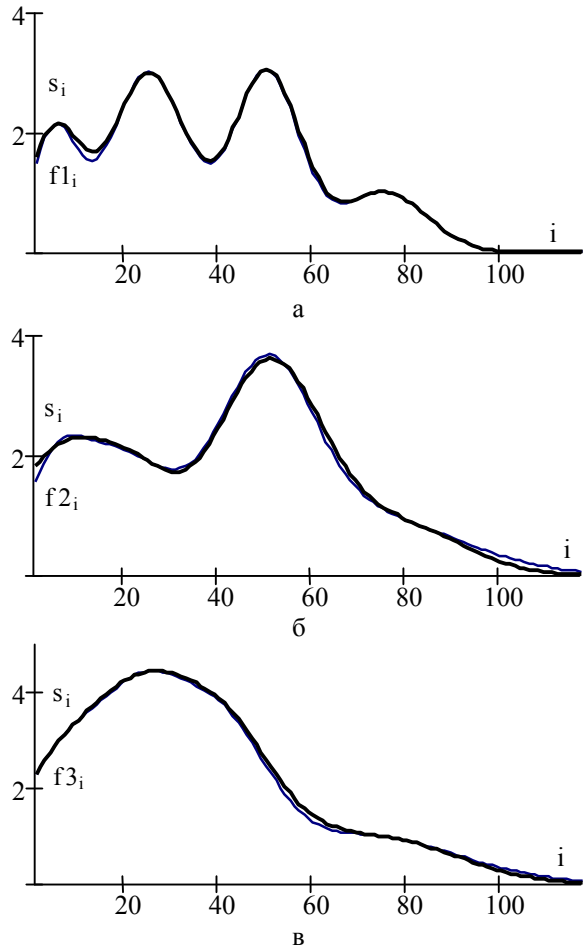

Рис. 1. Графики тестовых массивов и результаты их аппроксимации моделью (2):
а – количество локальных максимумов K^* равно количеству компонент смеси K, $K^* = K = 4$;
б – $K^* = 2$, $K = 4$; в – $K^* = 1$, $K = 4$

Для повышения точности оценок параметров рекомендуется увеличить количество итераций r. Как показано на рис. 2, при r порядка 60 средние относительные погрешности оценки параметров модели (2), аппроксимирующей массив $\{f3_i\}$, принимают значения: $\bar{\delta}_m < 10\ \%$, $\bar{\delta}_\sigma < 15\ \%$.

Поскольку зависимость критерия E_r (14) от количества итераций r не является монотонно убывающей (см. рис. 3), то в качестве условия остановки счета можно выбрать условие вида

$$(J_E \le \varepsilon_a) \wedge (r \ge N_r), \qquad (22)$$

где J_E – принятый критерий подобия объекта ($\{f_i\}$) и модели ($\{s_i\}$) (в частности, $J_E \equiv E_r$);

ε_a – допустимая погрешность;

r – номер итерации;

N_r – заданное минимальное число итераций.

В тех случаях, когда количество компонент K априори задано (например, в задачах распознавания для описания многомодальных эмпирических распределений при известном количестве классов объектов) или когда оценка K на двух последовательных шагах (r) и (r – 1) итерационной процедуры не изменяется, то для остановки счета можно использовать критерий, основанный на какой-либо мере расстояния между векторами оценок параметров модели, полученных на этих шагах

$$\max_{1\le\upsilon\le 3}\left\{\rho\left(\vec{\Theta}_\upsilon^{(r-1)}, \vec{\Theta}_\upsilon^{(r)}\right)\right\} \le \varepsilon_a, \qquad (23)$$

где $\vec{\Theta}_\upsilon$ – υ-й вектор множества параметров модели, $\left\{\vec{\Theta}\right\} = \left\{\vec{A}, \vec{m}, \vec{\sigma}\right\}$;

$\rho(\bullet,\bullet)$ – принятая метрика, например:

$$\max\left\{\left|\Theta_{k\upsilon}^{(r-1)} - \Theta_{k\upsilon}^{(r)}\right|\right\},\ \upsilon = 1\ldots 3,\ k = 1\ldots K;$$

$$\sum_{k=1}^{K}\left|\Theta_{k\upsilon}^{(r-1)} - \Theta_{k\upsilon}^{(r)}\right|,\ \upsilon = 1\ldots 3;$$

$$\left[\sum_{k=1}^{K}\left(\Theta_{k\upsilon}^{(r-1)} - \Theta_{k\upsilon}^{(r)}\right)^2\right]^{1/2},\ \upsilon = 1\ldots 3.$$

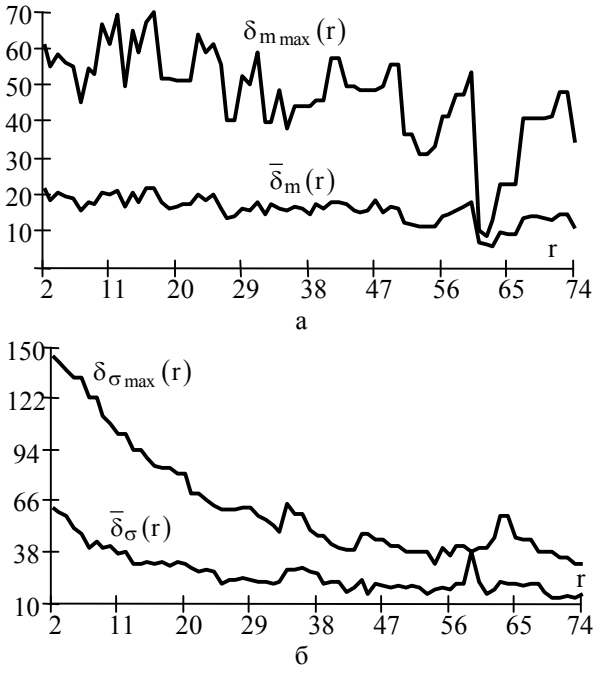

Рис. 2. Зависимость максимальных и средних относительных погрешностей оценок параметров смеси для описания $\{f3_i\}$ от количества итераций r:
а – m; б – σ

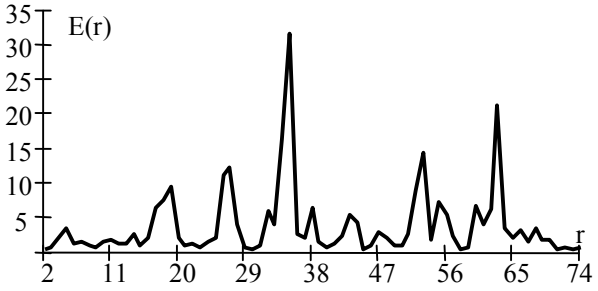

Рис. 3. Зависимость ошибки аппроксимации E от количества итераций r

На втором этапе апробации метода исследовалась его применимость для описания данных, структура зависимости которых отличается от принятой модели, в частности, для описания цифровых границ объектов. В качестве тестовой границы была взята 2d-форма частицы Чебышева – частично вогнутая окружность, которая в полярной координатной системе определяется выражением

$$p(\phi) = R \cdot [1 + \alpha \cos(n\phi)], \qquad (24)$$

где R – радиус недеформированной окружности;

α – параметр деформации, $|\alpha| < 1$;

n – степень полинома Чебышева при n = 3, $\alpha = -0{,}3$.

Для представления цифровой границы в комплексном виде $x_i + jy_i$, $i = 0, \ldots, M-1$ (при равномерной дискретизации угла $\phi \in [0, 2\pi]$) были сформированы кортежи значений массивов $\langle x_i, y_i \rangle$, где $\{x_i\}$ и $\{y_i\}$ – развертки границы по осям в декартовой системе координат:

$$x_i = p(\phi_i) \cdot \cos(\phi_i) - \min\{x_i\};$$
$$y_i = p(\phi_i) \cdot \sin(\phi_i) - \min\{y_i\},$$

т. о., $\forall i : x_i \geq 0, y_i \geq 0$.

На рис. 4, а, б показаны графики разверток $\{x_i\}$, $\{y_i\}$ и их аппроксимации смесью (2) $\{X_i\}$, $\{Y_i\}$ при r = 5; значения критерия E_5, соответственно, равны 0,137 и 0,038, на рис. 5 совмещены вид исходной границы и график кривой, рассчитанной по модели (2).

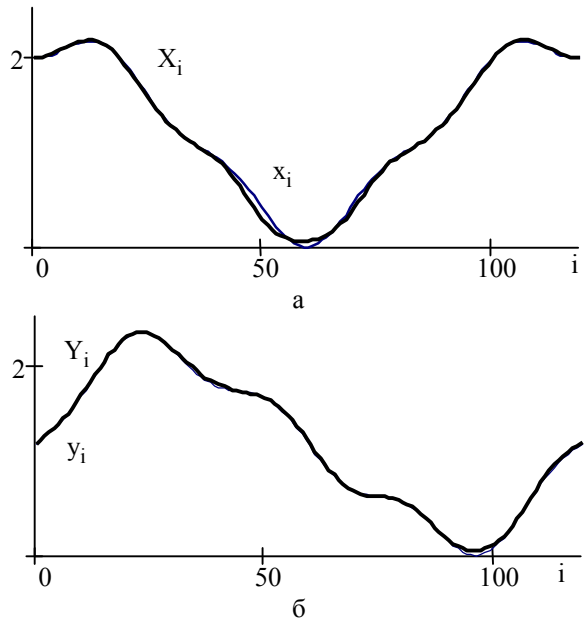

Рис. 4. Развертки границы по координатным осям и их представления моделью (2):
а, б – $\{x_i\}$ и $\{y_i\}$ – массивы значений координат x и y; $\{X_i\}$ и $\{Y_i\}$ – массивы результатов аппроксимации соответствующих разверток

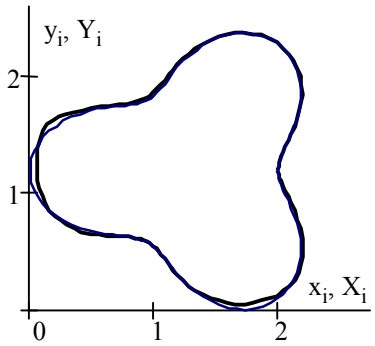

Рис. 5. Исходная граница $\{x_i, y_i\}$ и граница, восстановленная по модели (2) $\{X_i, Y_i\}$

Оценка количества компонент смеси: для $\{X_i\}$ $K_5 = 12$, для $\{Y_i\}$ $K_5 = 17$. Применение (18) – (21) при $\delta_m = 10$ % приводит к снижению оценки K до 10, однако точность описания формы кривых ухудшается: $E_5 = 0{,}504$ ($\{x_i\}$) и $E_5 = 1{,}013$ ($\{y_i\}$).

Заключение

Математическая модель в виде смеси базовых функций широко распространена на практике для описания неизвестных функциональных зависимостей, в частности, при параметрическом оценивании многомодальных эмпирических распределений. При этом по выборке наблюдений необходимо найти оценки для количества компонент смеси K, их удельных весов A_k (интерпретируемых как априорные вероятности классов) и параметров базовых функций (при выборе в качестве базовых функций Гаусса $N(z)$ – m_k и σ_k). Особенностью разработанного метода оценки параметров модели является отсутствие этапа предварительного статистического анализа описываемых данных с целью выбора оптимальных начальных приближений. Идея метода заключается в постепенном уточнении оценок параметров смеси по информации об ошибках аппроксимации на предыдущей итерации; основные расчетные формулы выведены из аналитической зависимости между точечными значениями функции $N(z)$ и ее производной $N'_z(z)$ с использованием конечно-разностных уравнений. При этом, если требуется описать вероятностное распределение, то полученные оценки весовых коэффициентов A_k следует переопределить как $A_k / (\sqrt{2\pi} \cdot \sigma_k)$, чтобы обеспечить условие нормировки базовых функций. Кроме того, метод позволяет достаточно точно аппроксимировать границу объектов, форма которых близка к умеренно деформированной окружности.

Т. о., данный метод может быть использован в задачах исследования геометрической и вероятностной природы совокупности анализируемых данных.

Литература

1. Гонсалес, Р. Цифровая обработка изображений [Текст] / Р. Гонсалес, Р. Вудс. – М. : Техносфера, 2005. – 1072 с.

2. Duda, R. O. Pattern Classification [Text] / R. O. Duda, P. E. Hart, D. G. Stork. – New York : John Wiley & Sons, 2001. – 654 p.

3. Гостев, И. М. Методы идентификации графических объектов на основе геометрической корреляции [Текст] / И. М. Гостев // Физика элементарных частиц и атомного ядра. – 2010. – Т. 41, Вып. 1. – С. 48 – 96.

4. Kalmykov, V. Structural analysis of contours as the sequences of the digital straight segments and of the digital curve arcs [Electronic resource] / V. Kalmykov // International Journal "Information Theories & Applications". – 2007. – Vol. 14. – P. 237 – 242. – Access mode: http://www.foibg.com/ijita/vol14/ijita14-3-p07.pdf. – 16.05.2015.

5. Фукунага, К. Введение в статистическую теорию распознавания образов [Текст] : пер. с англ. / К. Фукунага. – М. : Наука, 1979. – 368 с.

6. Попов, А. В. Автоматическая классификация данных дистанционного зондирования на основе теоретико-информационных критериев [Текст] / А. В. Попов, А. Н. Брашеван // Радіоелектронні і комп'ютерні системи. – 2009. – № 2 (36). – С. 120 – 129.

7. Прикладная статистика: классификация и снижение размерности [Текст] / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин ; под ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 608 с.

8. *Ветров, Д. П. Автоматическое определение количества компонент в EM-алгоритме восстановления смеси нормальных распределений [Текст] / Д. П. Ветров, Д. А. Кропотов, А. А. Осокин // Журнал вычисл. матем. и матем. физ. – 2010. – Т. 50, № 4. – С. 770 – 783.*

Поступила в редакцию 3.09.2015, рассмотрена на редколлегии 11.09.2015

ІТЕРАЦІЙНИЙ МЕТОД ОЦІНКИ ПАРАМЕТРІВ СУМІШІ ФУНКЦІЙ ГАУСА В ЗАДАЧАХ ОПИСУ ДАНИХ СПОСТЕРЕЖЕНЬ

І. К. Васильєва

Запропоновано математичну модель у виді суміші базових функцій для опису форм кривих, заданих масивами вибірок. Описано методику оцінки параметрів моделі, що складається в послідовному уточненні множин векторів оцінок параметрів, отриманих на основі інформації про помилки апроксимації на кожнім кроці ітераційної процедури, і надано основні розрахункові співвідношення. Приведено результати дослідження точності опису контрольних масивів із різною структурою взаємозв'язку даних сумішами функцій Гауса із використанням запропонованого методу оцінки параметрів моделі. Показано, що даний метод є ефективним і може застосовуватися для визначення кількості компонент суміші функцій Гауса й оцінок параметрів цих компонентів для опису масивів даних у задачах розпізнавання образів і кластерного аналізу.

Ключові слова: розпізнавання образів, опис форми, границя образа, параметрично задана крива, апроксимація, суміш базових функцій, оцінка параметрів суміші, ітерація, критерій точності.

AN ITERATIVE METHOD FOR ESTIMATING THE PARAMETERS OF A MIXTURE OF GAUSSIAN FUNCTIONS IN PROBLEMS OF THE DESCRIPTION OF OBSERVATIONAL DATA

I. K. Vasilyeva

The mathematical model having the form of basic functions mixture for description of the curves shapes defined by arrays of samples is offered. A method for estimating the model's parameters is described, which consists of the sequential specification of the sets of parameters estimations vectors obtained on the basis of information regarding approximation error at each step of the iterative procedure, and the basic relations for the calculations are given. The results of the investigation of accuracy of the description by the mixtures of Gaussian functions using the proposed method to estimate model parameters are represented for control arrays having different structures of the relationship among data. It is shown that this method is effective and can be used to determine the number of components of a mixture of Gaussian functions as well as parameter estimates of these components to obtain the descriptions of data arrays in problems of pattern recognition and cluster analysis.

Key words: pattern recognition, shape description, an object boundary, a parametrically defined curve, approximation, a mixture of the basic functions, estimation of mixture parameters, iteration, accuracy criterion.

Васильева Ирина Карловна – канд. техн. наук, доцент, доцент кафедры производства радиоэлектронных систем летательных аппаратов, Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: i.vasilyeva@mail.ru.