ISSN 1727-7337 (print)
ISSN 2663-2217 (online)

УДК 528.854

doi: 10.32620/aktt.2019.3.08

И. К. ВАСИЛЬЕВА, А. В. ПОПОВ

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

МЕТОД АВТОМАТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ

Предметом *изучения в статье являются методы автоматической кластеризации данных дистанционного зондирования в условиях априорной неопределенности относительно количества классов наблюдаемых объектов и статистических характеристик признаков классов.* ***Целью*** *является разработка метода аппроксимации многомодальных эмпирических распределений данных наблюдений для построения решающих правил в процедурах попиксельной статистической классификации, а также исследование эффективности данного метода при автоматической классификации объектов на синтезированных и реальных изображениях.* ***Задачи****: разработать и реализовать процедуру расщепления смеси базисных распределений, обеспечив при этом следующие требования: отсутствие этапа предварительного анализа данных с целью выбора оптимальных начальных приближений; хорошую сходимость метода и возможность автоматического уточнения списка классов путем объединения в один кластер неразличимых или слабо различимых компонент смеси; синтезировать тестовые изображения с заданным количеством объектов и известными законами распределения данных для каждого объекта; оценить эффективность разработанного метода автоматической классификации по критерию вероятности правильного распознавания; оценить результаты автоматической кластеризации реальных изображений. Используемыми* ***методами*** *являются: методы стохастического моделирования, методы аппроксимации эмпирических распределений, статистические методы распознавания, методы теории вероятностей и математической статистики. Получены следующие* ***результаты****. Предложен метод автоматического расщепления смеси распределений Гаусса для построения порогов принятия решения по критерию максимума апостериорной вероятности. Приведены результаты автоматического формирования списка классов и их вероятностных описаний, а также результаты кластеризации тестовых изображений и спутникового снимка. Показано, что разработанный метод является достаточно эффективным и может применяться для определения количества классов объектов и математического описания их стохастических характеристик в задачах распознавания образов и кластерного анализа.* ***Выводы.*** *Научная новизна полученных результатов состоит в том, что предлагаемый подход позволяет непосредственно в рамках процедуры обучения «без учителя» оценивать различимость классов и исключать неразличимые объекты из списка классов.*

Ключевые слова: *распознавание образов; кластеризация; аппроксимация; смесь базисных функций; оценка параметров смеси; вероятность правильного распознавания.*

Введение

Системы дистанционного зондирования с аэрокосмических носителей широко применяются при решении задач экологического мониторинга окружающей среды, картографирования, предупреждения чрезвычайных ситуаций и т.д. Спецификой данных задач является, как правило, отсутствие полной и достоверной априорной информации о количестве, типе и спектральных характеристиках наблюдаемых объектов. В этом случае начальным этапом анализа съемочных данных является кластеризация – разделение пикселов изображения на группы (кластеры); критерием отнесения пикселов к тому или другому кластеру обычно служит подобие спектральных характеристик. В результате автоматически формируется список классов из статистически однородных групп пикселов. Принятие решения о принадлежности пиксела тому или иному классу связано с построением оценок функций правдоподобия [1]:

$$\hat{L}_{kj} = \hat{f}_n(\vec{x}|a_k) \big/ \hat{f}_n(\vec{x}|a_j),$$

где $\vec{x} = \{x_1, \ldots, x_n\}$ – контрольная выборка (измеренные значения сигнальных признаков в текущем пикселе изображения);

$\hat{f}_n(\vec{x}|a_*)$ – оценка условной n-мерной плотности распределения вероятностей (ПРВ) выборочных значений признака $\vec{x}$ при условии принадлежности выборки к классу a_*.

В решающем правиле оценка отношения правдоподобия сравнивается с некоторым порогом, значение которого определяется выбранным критерием качества. Обучение классификатора сводится к задаче непараметрического оценивания ПРВ: по N точкам n-мерного пространства измерений $X = \{x_{ij}\}_{N \times n}$

нужно восстановить вид K поверхностей $f_n(\vec{x}|a_k)$ (k = 1…K) в (n + 1)- мерном пространстве, причем значение K заранее неизвестно. Одной из распространенных математических моделей, используемых для описания многомодальных ПРВ в задачах автоматической классификации «без учителя» является смесь базисных функций. Каждый класс интерпретируется как одномодальная генеральная совокупность (при неизвестном значении определяющего ее параметра), а классифицируемые наблюдения – как выборка из смеси таких совокупностей; при этом число компонент смеси определяет количество классов K, а удельных веса этих компонент – их априорные вероятности. В качестве базисных функций обычно выбирают нормальные распределения [2].

Модель в виде смеси гауссовых распределений (Gaussian Mixture Model – GMM) широко используется для решения различных задач, связанных с обработкой и анализом информации, в частности, для шумоподавления [3], сжатия изображений [4], сегментации оптических и радиолокационных снимков [5, 6], распознавания эмоций [7], рукописного текста [8] и т.п. Однако алгоритмы построения GMM и методика применения этой модели существенно зависят от характера исходных данных и специфики решаемой задачи [9]. Расщепление смеси распределений, как правило, выполняется в рамках одной из двух логических схем. В первой реализуется логика «от оценивания параметров смеси к классификации» (напр., EM-алгоритмы [2], основанные на методе максимального правдоподобия или методе моментов). Недостатками этого подхода являются значительный объём вычислений на этапе обучения и трудности при выборе начальных значений параметров. Во второй схеме, наоборот, 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 [2]).

В [10] предложен итерационный метод оценки параметров смеси ненормированных функций Гаусса N(x) для описания геометрической формы объектов. Особенностью разработанного является отсутствие этапа предварительного статистического анализа данных с целью выбора оптимальных начальных приближений, быстрая сходимость и регулируемая точность аппроксимации условных распределений. Идея метода заключается в постепенном уточнении оценок параметров смеси по информации о невязках аппроксимации на предыдущей итерации; основные расчетные формулы выведены из аналитической зависимости между точечными оценками параметров функции N(x) и ее производной с использованием конечно-разностных уравнений.

В данной работе описана методика адаптации данного метода для автоматической кластеризации данных дистанционного зондирования (ДЗ) и приведены результаты неконтролируемой классификации тестовых и реальных изображений.

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

1. Описание метода

Предлагаемая процедура автоматической кластеризации изображений состоит из нескольких этапов. На первом этапе выполняется преобразование исходного признакового пространства X в пространство меньшей размерности Y. Вид преобразования определяется спецификой данных наблюдений.

Так, при использовании в решающих правилах геометрических атрибутов образов, двумерную форму объектов, как правило, представляют с помощью некоторой одномерной функции $g(x)$. Примерами таких описаний являются функция тангенциального представления угла, противоположного отрезку дуги границы; комплексная функция $\dot{g}(x,y) = x + jy$, где $\langle x, y \rangle$ – развертки границы объекта по осям в декартовой системе координат, помещенной в центр тяжести образа объекта. [11]. Если процедура распознавания основывается на спектральных признаках, то обычно учитывают статистику распределения яркости пикселей изображения. При наличии цветных изображений целесообразно учитывать информацию о цвете пикселей. Для этого удобно использовать канал цветового тона (Hue) при представлении изображений в пространстве HSV (hue-saturation-value). Этот канал характеризует цвет точки, но не зависит от ее яркости, и, следовательно, является инвариантным к разностям фотометрических характеристик изображений. При анализе результатов многоспектральных дистанционных исследований можно использовать метод главных компонент, суть которого состоит в декоррелирующем преобразовании исходного пространства признаков и отборе спектральных компонент, соответствующих трем наибольшим собственным числам корреляционной матрицы. Три главные компоненты интерпретируются как три основных цвета (RGB), после чего выполняется их пересчет в значения Hue.

Следующий этап – построение гистограммы признака $\{f_i\}$, j = 0,M – 1; при 256 уровнях градации значений рекомендуемое количество интервалов гистограммы M = (40…50), что позволяет сгладить резкие выбросы эмпирического распределения и при этом сохранить характерный вид его огибающей.

ISSN 1727-7337 (print)
ISSN 2663-2217 (online)

В качестве базисных функций смеси, аппроксимирующей гистограмму, предлагается использовать ненормированные функции Гаусса

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

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:

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

где K – количество компонент смеси;

A_k – весовой коэффициент k-й компоненты;

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

Оценке подлежат следующие параметры смеси: $\left\{K,\vec{A},\vec{m},\vec{\sigma}\right\}$. Для нахождения значений m и σ k-й компоненты смеси (нижние индексы опущены) нужно решить дифференциальное уравнение:

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

Заменив в (3) производную конечно-разностным выражением

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

получим:

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

i = 1, M – 2.

Корни системы (4):

$$\hat{m}_i=\left(i+\frac{f_{i+1}^2-f_{i+1}\cdot f_{i-1}}{-f_{i+2}\cdot f_i+f_{i+1}^2-f_{i+1}\cdot f_{i-1}+f_i^2}\right)\cdot\Delta z; \quad (6)$$

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

Расчеты по (6), (7) следует проводить только для подмножества индексов элементов $\{f_i\}$ $J\subset i$, соответствующих точкам локальных максимумов, а также для граничных точек массива, если те являются верхними гранями своих малых окрестностей.

Поскольку положение локальных максимумов локализуется с точностью до 0,5Δz (где Δz – ширина интервала гистограммы), то погрешность оценок m_k, σ_k зависит от разрядности гистограммы K (при Δz = const и фиксированном размахе случайной величины z; если для динамического диапазона яркости используют восьмибитное представление, то $z\in[0, 255]$). Кроме того, точность оценок параметров распределений смеси (2) зависит от погрешностей, обусловленных конечно-разностным представлением производной (4) и влиянием остальных компонент смеси.

Мощность подмножества индексов {J} является текущей оценкой количества компонент смеси K.

Поскольку $A_J\cdot N(m_J;m_J,\sigma_J)=A_J$, то предварительные оценки весовых коэффициентов определяются значениями локальных максимумов (мод эмпирического распределения признака z)

$$A_J^*=f_J. \quad (8)$$

Уточненные оценки коэффициентов A_J находят по формуле:

$$A_J=A_J^*-\sum_{k,k\neq J}A_k\cdot N(m_J;m_k,\sigma_k). \quad (9)$$

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

По значениям модели (2) при текущих значениях параметров $s_i^{(r)}$ можно провести валидацию модели – оценить ее точность по некоторому критерию, например, найти среднеквадратическую ошибку (mean square error – MSE):

$$E_r=\frac{1}{M}\sum_{i=1}^{M}\left(f_i-s_i^{(r)}\right)^2. \quad (10)$$

Если точность описания гистограммы моделью (2) оказывается недостаточной, то формируют разностный массив $\{d_i\}$:

$$d_i=f_i-s_i^{(r)}; \quad (11)$$

при этом если $|d_j|<\varepsilon_d$, где ε_d – заданный порог, то соответствующее значение d_j, j = 1,…,M считается незначимым и обнуляется.

Если в разностном массиве все значимые по величине элементы неотрицательны, то множество $\{d_i\}$ является дополнением к множеству значений модели (2) $\left\{s_i^{(r)}\right\}$, полученному на текущем этапе; в этом случае массив $\{d_i\}$ заменяет первоначальный массив $\{f_i\}$, и итерационная процедура оценки параметров смеси продолжается до тех пор, пока не будет достигнута приемлемая точность описания данных моделью (2); при этом множества оценок параметров на шагах r и (r – 1) объединяются:

$$\Theta^{(r)}=\Theta^{(r)}\cup\Theta^{(r-1)},$$

где $\Theta\subset\left\{\vec{m},\vec{\sigma},\vec{A}\right\}$;

$$K_r=K_r+K_{r-1}.$$

Дополнительным критерием остановки счета может служить превышение заданного количества итераций.

Если в разностном массиве некоторые из элементов – отрицательные, то текущие оценки параметра σ переопределяют по эмпирической формуле:

$$\text{если } \sigma_k > \xi, \text{ то } \sigma_k = \xi; \qquad (12)$$

$$\xi = \frac{|\mu - m_k|}{\sqrt{|2\ln(\varepsilon/|A_k|)|}},$$

где μ – индекс элемента $d_{min} = \min\{d_i\}$;

ε – поправочный коэффициент, 0 < ε < 1; значение ε выбирается таким, чтобы обеспечить выполнение условия $\forall i : d_i \geq 0$.

Если количество компонент K априори задано (например, в задачах распознавания для описания многомодальных эмпирических распределений при известном количестве классов объектов), а оценка K, полученная на шаге r, $K_r > K$, то ближайшие (по значению параметра m) компоненты объединяются

$$m_\eta^{(r)} = \frac{A_\nu^{(r-1)} m_\nu^{(r-1)} + A_\eta^{(r)} m_\eta^{(r)}}{A_\nu^{(r-1)} + A_\eta^{(r)}}, \qquad (13)$$

$$\sigma_\eta^{(r)} = \sqrt{\frac{A_\eta^{(r)} \sigma_\eta^{(r)2}}{A_\eta^{(r)} + A_\nu^{(r-1)}} + \frac{A_\nu^{(r-1)} \sigma_\nu^{(r-1)2}}{A_\eta^{(r)} + A_\nu^{(r-1)}}}, \qquad (14)$$

где $\nu \in \{J\}^{(r-1)}$, $\eta \in \{J\}^{(r)}$,

после чего по (8), (9) пересчитываются оценки $A_\eta^{(r)}$.

Если количество классов наблюдаемых объектов заранее не известно, а оценка K на двух последовательных шагах итерационной процедуры не изменяется, то для остановки счета можно использовать критерий, основанный на какой-либо мере расстояния между векторами оценок параметров модели, полученных на шагах r и (r – 1):

$$\max_{1 \leq \upsilon \leq 3} \left\{ \rho\left(\Theta_\upsilon^{(r-1)}, \Theta_\upsilon^{(r)}\right) \right\} \leq \varepsilon_a,$$

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

$$\max \left\{ \left| \Theta_{k\upsilon}^{(r-1)} - \Theta_{k\upsilon}^{(r)} \right| \right\}, \ \upsilon = 1\ldots3, \ k = 1\ldots K;$$

$$\sum_{k=1}^{K} \left| \Theta_{k\upsilon}^{(r-1)} - \Theta_{k\upsilon}^{(r)} \right|, \ \upsilon = 1\ldots3;$$

$$\left[\sum_{k=1}^{K} \left(\Theta_{k\upsilon}^{(r-1)} - \Theta_{k\upsilon}^{(r)} \right)^2 \right]^{1/2}, \ \upsilon = 1\ldots3.$$

В данной работе для уточнения количества компонент смеси предложен метод, основанный на оценке различимости классов по критерию максимума апостериорной вероятности; в качестве оценок априорных вероятностей классов принимаются весовые коэффициенты модели (2). Для нахождения порогов принятия решения (границ между парами соседних классов – кластеров X_i и X_{i+1}) массивы параметров модели (2) упорядочиваются так, чтобы $\forall i \in [1, K-1]$ выполнялось условие: $m_i < m_{i+1}$.

Искомые пороги $\{t_{1,i}, t_{2,i}\}$, i = 1…K –1 являются корнями системы квадратных уравнений

$$\ln\left[\frac{A_i}{\sqrt{2\pi}\sigma_i} N(t_i; m_i, \sigma_i) \right] = \ln\left[\frac{A_{i+1}}{\sqrt{2\pi}\sigma_{i+1}} N(t_i; m_{i+1}, \sigma_{i+1}) \right]. \qquad (15)$$

По значениям порогов $\{t_{1,i}\}$ можно уточнить состав компонент смеси. Признаками избыточности множества классов являются следующие: некоторые из элементов множества $\{t_{1,i}\}$ имеют комплексные значения; последовательность значений элементов $\{t_{1,i}\}$ не упорядочена по возрастанию. В любом случае, если $t_{1,u}$ – комплексное число или если $t_{1,u} > t_{1,u+1}$ или $t_{1,u} < t_{1,u-1}$, то класс X_u исключается из списка классов. Также целесообразно исключить классы, для которых вероятность ошибочного распознавания практически равна единице (с точностью до заданной величины ε_p). Для двух соседних (упорядоченных по номерам) классов X_i и X_{i+1} вероятности ошибок $P_{i,i+1}$ и $P_{i+1,i}$, где первый индекс соответствует номеру класса, в пользу которого принято решение, а второй – действительному номеру класса, определяются по формулам:

$$P_{i,i+1} = \frac{1}{\sqrt{2\pi}\sigma_{i+1}} \left[\int_0^{t_{1,i}} N(z; m_{i+1}, \sigma_{i+1}) dz + \int_{t2_i}^{255} N(z; m_{i+1}, \sigma_{i+1}) dz \right]; \qquad (16)$$

$$P_{i+1,i} = \frac{1}{\sqrt{2\pi}\sigma_i} \int_{t_{1,i}}^{t_{2,i}} N(z; m_i, \sigma_i) dz. \qquad (17)$$

Если любая из этих вероятностей ошибок близка к единице, то классы X_u и X_{u+1} считаются неразличимыми и объединяются в один кластер.

Достоинством разработанного метода является то, что при уменьшении количества классов нет необходимости пересчитывать параметры модели (2), так как пороги принятия решения уже найдены – достаточно лишь отбросить из множества порогов те, которые ранее были границами между объединенными классами. Например, если объединяются классы X_u и X_{u+1}, то исключаются $t_{1,u}$ и $t_{2,u}$.

После исключения неразличимых классов, решение принимается в пользу того класса, чья апостериорная вероятность максимальна; процедура классификации состоит в определении принадлежности текущего пиксела изображения определенной обла-

ISSN 1727-7337 (print)
ISSN 2663-2217 (online)

сти принятия решения, ограниченной порогами $\{t^*_{1,i}, t^*_{2,i}\}$, $i = 1 \ldots K^* - 2$, где K^* – уточненное количество классов.

2. Результаты оценки эффективности метода автоматической кластеризации

В качестве контрольных данных на первом этапе апробации предлагаемого метода автоматической кластеризации были синтезированы два тестовых изображения (ТИ) с известным количеством классов объектов (K = 4). Изображения имели одинаковую пространственную структуру, соответствующую эталону (рис. 1), однако отличались параметрами распределения яркостных признаков классов и их априорными вероятностями $P(a_k)$; величина $P(a_k)$ интерпретировалась как относительная суммарная площадь образов объектов класса a_k. Формирование матриц яркостей ТИ № 1, ТИ № 2 выполнялось по блочному принципу: [**M**]: 4×4, где $M_{i,j}$ = [$\mathbf{B^k}$]: 50×50; т.о., каждый блок $\mathbf{B^k}$ представлял образ соответствующего класса a_k, суммарное количество блоков $\mathbf{B^k}$ определялось пропорционально $P(a_k)$, а значения элементов $B^k_{\mu,\nu}$ моделировались как случайные величины с гауссовым распределением и параметрами m_k и σ_k. Таким образом, модель (2) априори полностью соответствовала структуре взаимосвязи данных в контрольных выборках. Для отображения взаимных корреляционных связей между строками и столбцами изображений, сгенерированные массивы случайных яркостей в блоках $\mathbf{B^k}$ упорядочивались по значению и конвертировались в матрицы. На рис. 2, 3 представлены ТИ № 1, ТИ № 2, сформированные из [$\mathbf{B^k}$].

В качестве признакового описания объектов на ТИ был принят набор значений функций яркости пикселов образа соответствующего класса.

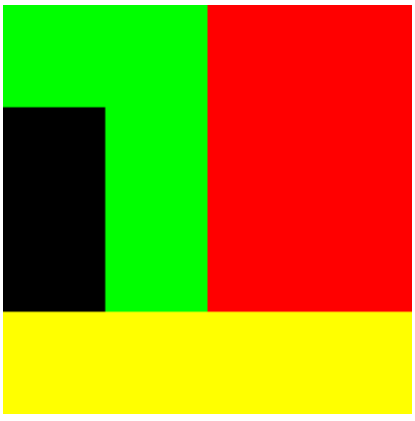


Рис. 1. Эталонное изображение

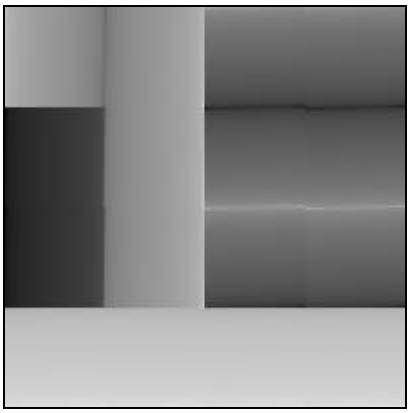


Рис. 2. Тестовое изображение ТИ № 1

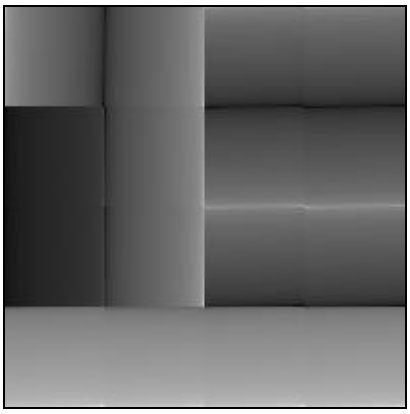


Рис. 3. Тестовое изображение ТИ № 2

Теоретические $s1^*(I)$, $s2^*(I)$ и эмпирические $f1_i$, $f2_i$ распределения яркостного признака I для ТИ № 1 и ТИ № 2, соответственно, показаны на рис. 4.

Результаты третьей итерации процедуры оценки параметров смеси распределений (2), представленных гистограммами (см. рис. 4, 5), приведены на рис. 6, 7 и в табл. 1; в табл. 1 также указаны фактические значения параметров смеси $\vec{m}$, $\vec{\sigma}$, $\vec{A}$ для ТИ № 1, ТИ № 2 и значения критерия точности модели E_3 (10). Для повышения точности оценивания высоты гистограмм брались с масштабирующим коэффициентом, равным 100.

Таблица 1

Значения параметров модели для тестовых массивов и их оценки (количество итераций r = 3)

$\vec{m}_1$	$\vec{m}_1^{(3)}$	$\vec{\sigma}_1$	$\vec{\sigma}_1^{(3)}$	$\vec{A}_1$	$\vec{A}_1^{(3)}$	E_3
50	51,58	12	12,05	0,125	0,154	0,048
100	98,51	20	18,63	0,375	0,346	
150	148,05	15	15,41	0,25	0,235	
200	200,10	10	10,85	0,25	0,265	
$\vec{m}_2$	$\vec{m}_2^{(3)}$	$\vec{\sigma}_2$	$\vec{\sigma}_2^{(3)}$	$\vec{A}_2$	$\vec{A}_2^{(3)}$	E_3
40	41,94	10	11,72	0,125	0,221	0,047
85	69,27	22	10,23	0,375	0,171	
100	94,91	25	17,55	0,25	0,391	
150	145,88	20	22,95	0,25	0,215	

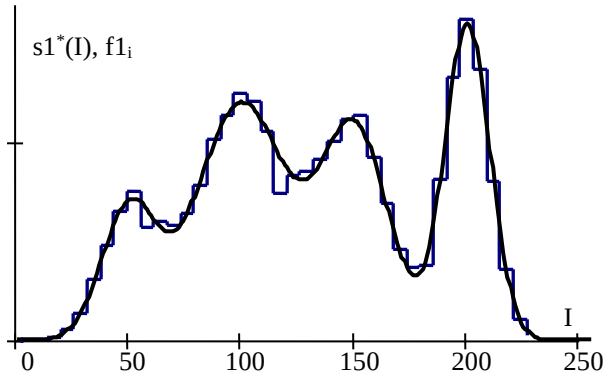


Рис. 4. Вид модели (2) и гистограммы распределения яркости пикселей ТИ № 1

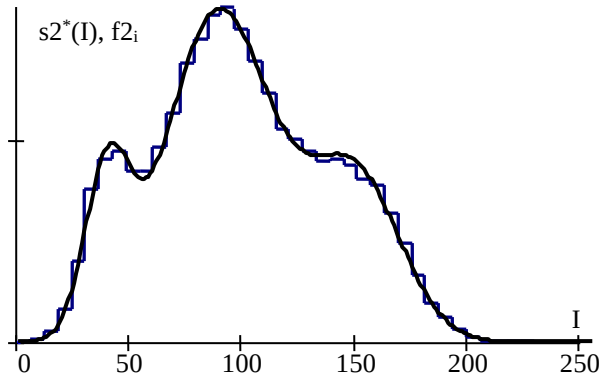


Рис. 5. Вид модели (2) и гистограммы распределения яркости пикселей ТИ № 2

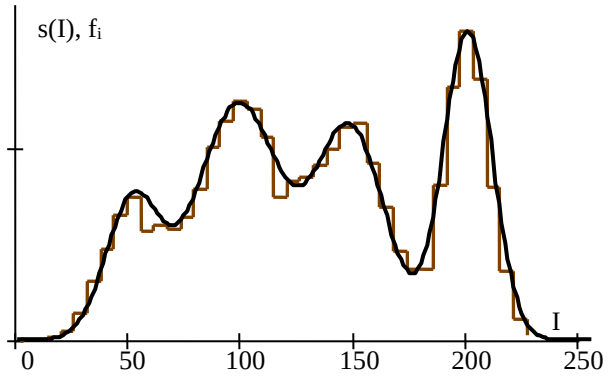


Рис. 6. Результаты оценки параметров модели (2) для ТИ № 1

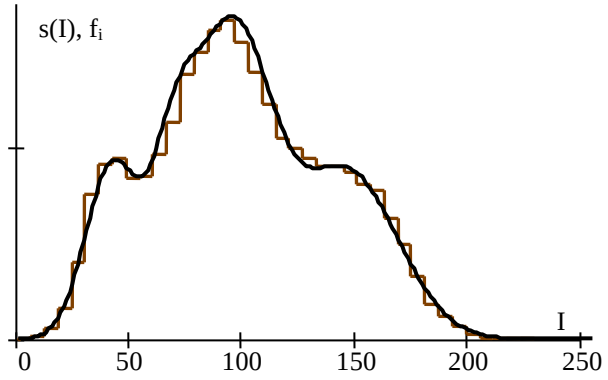


Рис. 7. Результаты оценки параметров модели (2) для ТИ № 2

Анализ результатов, представленных в табл. 1, показывает, что если классы достаточно разделимы (количество локальных максимумов функции ПРВ равно числу компонент смеси), то метод позволяет уже за два – три шага достигнуть приемлемой точности как описания формы кривой ПРВ, так и определения значений параметров смеси. Так, максимальные относительные погрешности оценок параметров модели (2) m и σ для ТИ № 1 составляют, соответственно, $\max\{\delta_m\} < 3{,}5\ \%$, $\max\{\delta_\sigma\} < 9\ \%$.

Для слабо разделимых классов даже при удовлетворительном совпадении графиков контрольных массивов и графиков результатов их аппроксимации моделью (2) оценки параметров смеси могут существенно отличаться от их фактических значений. Так, значения критерия E_3, характеризующего точность аппроксимации моделью (2) гистограмм яркости ТИ № 1 и ТИ № 2 практически одинаковы: 0,048 и 0,047, соответственно. При этом относительные погрешности оценок параметров смеси (2) m и σ для ТИ № 2 значительно больше, чем для ТИ № 1: $\max\{\delta_m\} < 20\ \%$, $\max\{\delta_\sigma\} < 55\ \%$.

Поскольку все статистические решающие правила основаны на информации о распределении выборочных значений (при условии их принадлежности к определенному классу), то от точности оценки параметров условных ПРВ во многом зависит и эффективность классификации. С этой точки зрения, для оценки адекватности модели (2) целесообразно использовать не среднеквадратичную ошибку (10), а критерии достоверности результатов классификации; в качестве последних обычно используют значения вероятностей правильного распознавания (или их статистические оценки).

Матрица теоретических ошибок распознавания P_{ij} (16), (17) компонент смеси (2) для тестовых изображений представлена в табл. 2; диагональные элементы матрицы P_{ii} соответствуют вероятности правильной классификации для i-го класса.

Статистические оценки вероятностей ошибок P_{ij} были получены по результатам попиксельной классификации тестовых изображений по критерию максимума апостериорной вероятности, в соответствии с которым контрольная выборка (значения признаков в текущем пикселе ТИ) относилась к тому классу a_n, $1 \le n \le K$, апостериорная вероятность которого превышала апостериорные вероятности остальных классов:

$$f(z|a_n) = \max_{1 \le k \le K} \{P(a_k) f(z|a_k)\} \Rightarrow z \in a_n,$$

где $f(z|a_k)$ – ПРВ (1) k-й компоненты модели (2).

Для определения порогов принятия решения по (15) использовались два подхода: обучение «с учителем» – при априори известных параметрах услов-

ISSN 1727-7337 (print)
ISSN 2663-2217 (online)

ных по классам ПРВ $f(z|a_k)$, и обучение «без учителя» – параметры $f(z|a_k)$ были найдены описанным выше методом. Таким образом, сравнительный анализ показателей достоверности классификации при известных эталонных описаниях классов и при описаниях, полученных путем расщепления смеси распределений, позволил оценить эффективность предлагаемого метода расщепления.

Результаты оценки достоверности классификации объектов на ТИ № 1, ТИ № 2 для обучения «с учителем» и «без учителя» приведены в табл. 3, 4, где через γ_i обозначены решения в пользу класса a_i: $P_{ij} = P\{\gamma_i|a_j\}$. Оценка общей вероятности правильной классификации $P_{пр}$ вычислялась с учетом априорных вероятностей классов $P(a_k)$ на тестовых изображениях:

$$P_{пр} = \sum_{k=1}^{K} P(a_k)P_{kk} .$$

Таблица 2

Матрица теоретических ошибок классификации P_{ij} объектов на тестовых изображениях

	ТИ № 1				ТИ № 2			
P_{ij}	1	2	3	4	1	2	3	4
1	0,91	0,04	0	0	0,89	0,07	0,03	0
2	0,09	0,89	0,08	0	0,11	0,79	0,64	0,02
3	0	0,07	0,89	0,02	0	0,11	0,18	0,09
4	0	0	0,03	0,98	0	0,03	0,15	0,89

Таблица 3

Статистические оценки вероятностей ошибок классификации P_{ij} объектов на ТИ № 1

	Обучение «с учителем»				Обучение «без учителя»			
P_{ij}	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
γ_1	0,90	0,04	0	0	0,95	0,06	0	0
γ_2	0,10	0,89	0,08	0	0,05	0,84	0,05	0
γ_3	0	0,07	0,89	0,02	0	0,10	0,91	0,01
γ_4	0	0	0,03	0,98	0	0	0,04	0,99
$P_{пр}$	0,91				0,91			

Таблица 4

Статистические оценки вероятностей ошибок классификации P_{ij} объектов на ТИ № 2

	Обучение «с учителем»				Обучение «без учителя»			
P_{ij}	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
γ_1	0,90	0,07	0,03	0	0,98	0,14	0,06	0
γ_2	0,10	0,80	0,62	0,02	0,02	0,19	0,10	0
γ_3	0	0,10	0,19	0,09	0	0,62	0,64	0,08
γ_4	0	0,03	0,16	0,89	0	0,05	0,20	0,92
$P_{пр}$	0,73				0,65			

Полученные результаты свидетельствуют о том, что при условии достаточной статистической разделимости классов в пространстве признаков (случай, представленный на ТИ № 1), данный метод позволяет обеспечить достоверность распознавания классов на таком же уровне, что и процедуры обучения «с учителем»: $P_{пр} = 0,91$. ТИ № 2 представляет случай слабой разделимости классов 2, 3. Так, даже для априори известных эталонных описаниях классов вероятность правильного распознавания класса 3 $P_{33} = 0,19$, а вероятность ошибки $P_{23} = 0,62$.

При автоматической классификации слабо различимых объектов резко возрастают погрешности оценки параметров компонент смеси (т.е., оценки условных по классам ПРВ, которые используются в решающих правилах), поэтому показатели достоверности распознавания классов снижаются: $P_{пр} = 0,65$ (при контролируемой классификации $P_{пр} = 0,73$). Стоит отметить, что ухудшение критерия качества $P_{пр}$ обусловлено слабо различимыми классами 2, 3 (см. рис. 8); при этом если при контролируемой классификации класс 3 в 62 % случаев распознавался как класс 2, то при автоматической классификации, наоборот, с вероятностью $P_{32} = 0,62$ пикселы класса 2 ошибочно помечались маркером класса 3. Ошибки такого рода могут быть скорректированы путем переопределения номеров классов.

Таким образом, по результатам первого этапа апробации предлагаемого метода на тестовых изображениях, можно сделать вывод об его эффективности для решения задачи отнесения объектов к некоторым классам, количество и свойства которых априори не известны.

Визуализация результатов статистической попиксельной классификации тестовых изображений (в виде карт классов) представлена на рис. 9, 10.

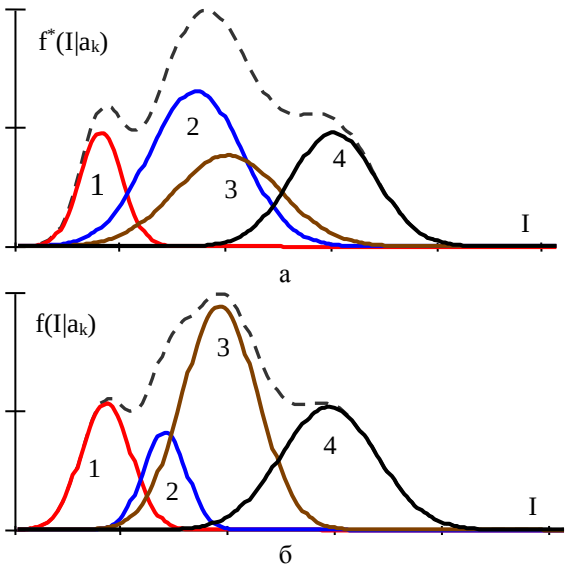
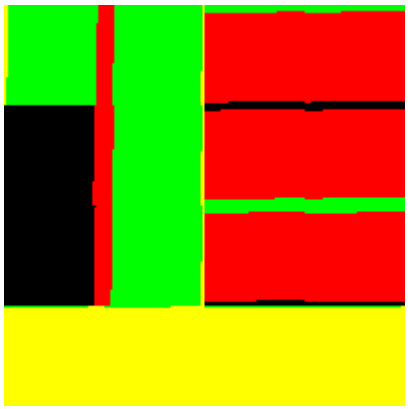
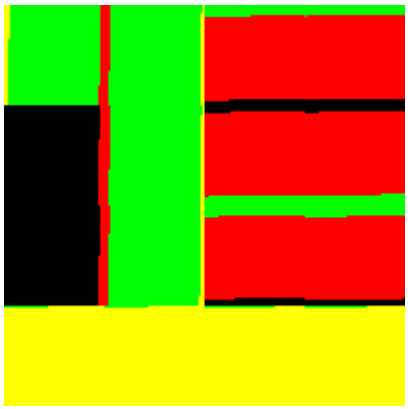


Рис. 8. ПРВ классов (а) и их оценки (б) для ТИ №2



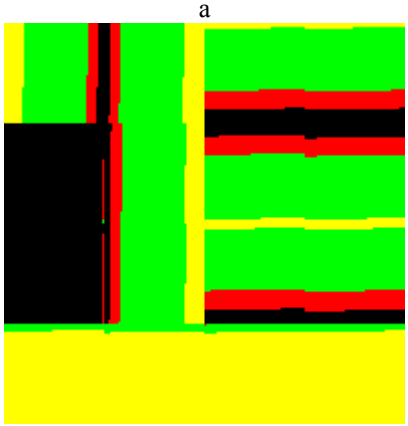
а



б

Рис. 9. Результаты классификации ТИ №1:
а – обучение «с учителем»; б – «без учителя»

а



б

Рис. 10. Результаты классификации ТИ №2:
а – обучение «с учителем»; б – «без учителя»

На втором этапе апробации с помощью предлагаемого метода была проведена процедура классификации спутникового снимка, представленного на рис. 11. Гистограмма распределения яркостного признака V (компоненты «value» цветовой модели HSV) показана на рис. 12.

Результаты оценки количества классов и параметров их распределений (в соответствии с принятой моделью (2)) приведены в табл. 5 и на рис. 13.

Как видно по данным, представленным в табл. 5, уже на второй итерации (r = 2) достигнута удовлетворительная точность аппроксимации гистограммы моделью (2), однако наличие среди множества весовых коэффициентов $\{A_k\}$ отрицательных значений обусловило необходимость продолжить итерации метода с тем, чтобы уточнить состав компонент смеси распределений (2).

Уменьшение числа компонент на следующих итерациях метода связано с объединением неразличимых классов (по критерию максимальной апостериорной вероятности) на основании анализа порогов принятия решения $\{t_{1,i}\}$, как рассмотрено ранее.

Рис. 11. Исходное изображение

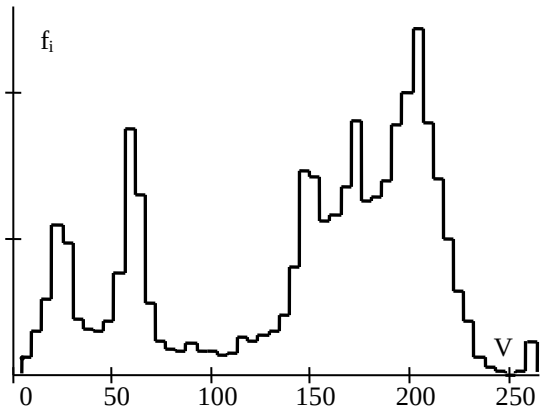


Рис. 12. Гистограмма признака V («value») исходного изображения (см. рис. 11)

ISSN 1727-7337 (print)
ISSN 2663-2217 (online)

Таблица 5

Результаты оценки числа классов и ошибки аппроксимации гистограммы моделью (2)

r	K_r	δ_{max}, %	δ_{min}, %	$\delta_{ср}$, %
1	9	75,7	0,4	23,2
2	21	46,5	4×10^{-4}	10,6
3	17	46,7	0,01	12,3
4	12	46,7	0,05	12,5
5	11	50,9	0,05	12,5

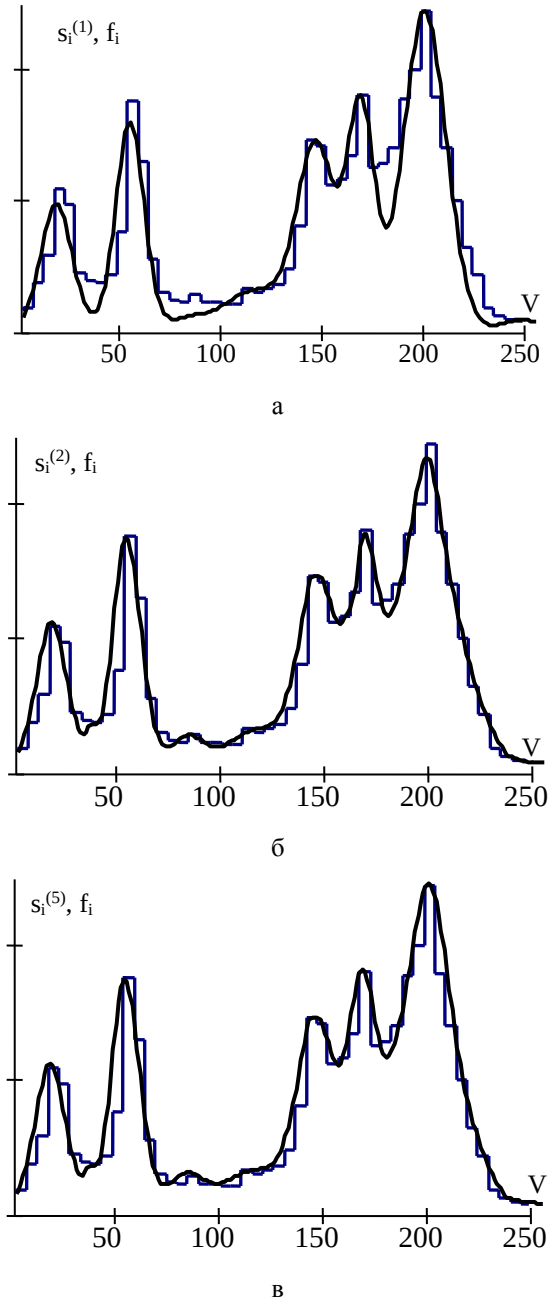


Рис. 13. Гистограмма распределения признака V и модель (2), полученная на r-й итерации процедуры расщепления смеси:
а – r = 1, K = 9; б – r = 2, K = 21; в – r = 5, K = 11

Для автоматической классификации реального изображения (см. рис. 11) были взяты оценки параметров смеси (2), полученные на пятой итерации метода; при r = 5 достигается компромисс между приемлемой точностью описания эмпирического распределения моделью (2) и количеством различимых классов объектов наблюдения. Дальнейшее уменьшение числа классов приводит к ухудшению описания формы гистограммы. Так, уже при K = 10 (r = 6) существенно возрастает относительная погрешность аппроксимации: максимальное значение δ_{max} = 81 %, среднее – $\delta_{ср}$ = 23 %.

Результаты автоматической классификации объектов на изображении (см. рис. 11) при K = 11 (что соответствует оптимальной оценке параметров компонент смеси распределений) и при K = 4 (что соответствует принятому ограничению на количество классов объектов) представлены на рис. 14, 15, соответственно.

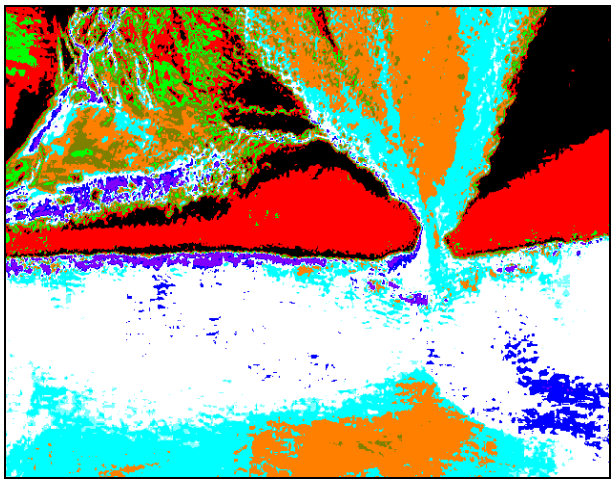


Рис. 14. Результаты автоматической классификации изображения (см. рис. 11) (K = 11)

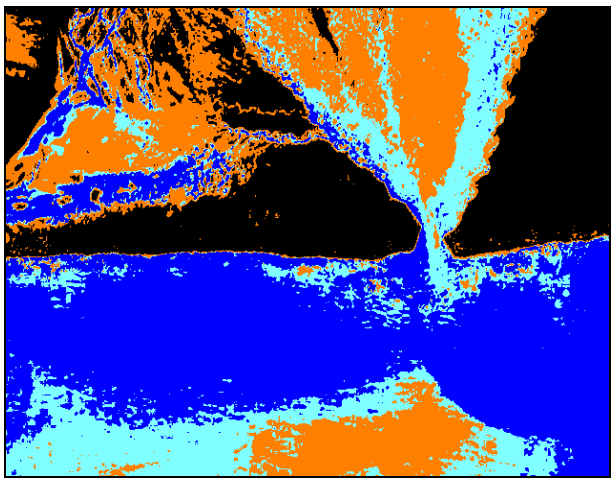


Рис. 15. Результаты автоматической классификации изображения (см. рис. 11) при ограничении числа классов (K = 4)

Рисунок 15 иллюстрирует результаты объединения нескольких классов, близких по статистическим характеристикам признаков распознавания (это классы с соседними номерами): $a_1^* = \{1, 2, 3\}$; $a_2^* = \{4, 5, 6\}$; $a_3^* = \{7, 8\}$; $a_4^* = \{9, 10, 11\}$.

Уменьшение количества классов может упростить процедуры последующего анализа и интерпретации результатов классификации. При этом следует особо отметить, что, несмотря на исходное предположение о нормальном распределении признаков объектов, результирующее распределение кластера, объединяющего несколько компонент, уже не будет являться гауссовым. Таким образом, требование соответствия реальных распределений признаков нормальному виду не является жестким условием и не ограничивает область применимости предлагаемого метода в процедурах автоматической классификации данных ДЗ, распределения которых зачастую имеют негауссов вид.

Заключение

Разработан метод автоматической классификации данных ДЗ в условиях априорной неопределенности относительно количества классов объектов и их статистических характеристик. Данный метод основан на представлении многомодального эмпирического распределения наблюдений в виде смеси ненормированных функций Гаусса. Оценки параметров базисных распределений, полученные в результате расщепления смеси, используются в качестве априорной информации при распознавании объектов по критерию максимума апостериорной вероятности. Уточнение списка классов выполняется путем исключения заведомо неразличимых объектов на основании оценок вероятностей ошибок распознавания; процедура исключения классов сводится к анализу множества порогов принятия решения, определяющих границы между классами. При объединении неразличимых (или слабо различимых) классов из текущего списка в один кластер, результирующее распределение кластера уже не будет являться гауссовым; этот факт позволяет обойти ограничение на требуемый вид распределения данных.

Предлагаемый метод позволяет автоматически выделять на изображении области со статистически однородными свойствами классификационных признаков. Дальнейшая обработка выделенных областей может включать морфологические методы для выявления топологических свойств объектов и описания основных закономерностей во вторичном структурном признаковом пространстве.

Как показали результаты апробации метода на синтезированных и реальных изображениях, предлагаемый подход позволяет эффективно решать задачи исследования вероятностной природы совокупности анализируемых данных, кластерного анализа и автоматизированной классификации данных ДЗ.

Прикладное значение результатов состоит в том, что предлагаемый подход позволяет в рамках процедуры обучения «без учителя» не только оценить неизвестное количество классов (компонентов смеси) и построить их статистические модели, но включить в автоматически формируемый список классов только реально различимые объекты на изображении. При этом разработанный итерационный метод расщепления смеси не требует оптимального выбора начальных приближений и обладает быстрой сходимостью.

Направление дальнейших исследований – разработка и исследование эффективности методов автоматической классификации объектов на многоканальных изображениях и использованием векторных признаков классов.

Литература

1. *Фукунага, К. Введение в статистическую теорию распознавания образов [Текст] : пер. с англ. / К. Фукунага. – М. : Наука, 1979. – 368 с.*
2. *Прикладная статистика: классификация и снижение размерности [Текст] / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1989. – 608 с.*
3. *Image Denoising Using Asymmetric Gaussian Mixture Models [Text] / W. He, R. Yu, Y. Zheng, T. Jiang // Internat. Symposium in Sensing and Instrumentation in IoT Era (ISSI), Shanghai. – 2018. – P. 1 – 4. doi: 10.1109/ISSI.2018.8538279.*
4. *Sun, J. Image Compression Using GMM Model Optimization [Text] / J. Sun, Y. Zhao, S. Wang // Acoustics, Speech and Signal Processing (ICASSP) : IEEE Proc. Internat. Conf., Brighton, United Kingdom, 2019. – Brighton, 2019. – P. 1797-1801. doi: 10.1109/ICASSP.2019.8683784.*
5. *Variational Bayesian Change Detection of Remote Sensing Images Based on Spatially Variant Gaussian Mixture Model and Separability Criterion [Text] / G. Yang, H. Li, W. Yang et al. // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing – 2019. – Vol. 12, no. 3. – P. 849 – 861. doi: 10.1109/JSTARS.2019.2896233.*
6. *Bouhlel, N. Unsupervised Segmentation of Multilook Polarimetric Synthetic Aperture Radar Images [Text] / N. Bouhlel, S. Méric // IEEE Transactions on Geoscience and Remote Sensing. – 2019. – P. 1-15. doi: 10.1109/TGRS.2019.2904401.*
7. *Shahin, I. Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network [Text] / I. Shahin, A. B. Nassif, S. Hamsa // IEEE Access. – 2019. – Vol. 7. – P. 26777-26787. doi: 10.1109/ACCESS.2019.2901352.*

ISSN 1727-7337 (print)
ISSN 2663-2217 (online)

8. Hollaus, F. MultiSpectral Image Binarization using GMMs [Text] / F. Hollaus, M. Diem, R. Sablatnig // Frontiers in Handwriting Recognition (ICFHR) : Proc. 16th Internat. Conf., Niagara Falls, NY, 2018. – 2018. – P. 570-575. doi: 10.1109/ICFHR-2018.2018 .00105.

9. Popov, A. V. Image clustering algorithm using polynormal distribution [Текст] / A. V. Popov, O. Pogrebnyak, A. N. Brashevan // Mathematical Methods in Pattern and Image Analysis : Proc. of SPIE. – 2005. – Vol. 5916. – P. 5916-5925.

10. Васильева, И. К. Итерационный метод оценки параметров смеси функций Гаусса в задачах описания данных наблюдений [Текст] / И. К. Васильева // Радіоелектронні і комп'ютерні системи. – 2015. – № 3 (73). – С. 70-76.

11. Vasil'eva, I. Multicomponent Model of Objects Attributive Signatures on Color Pictures [Text] / I. Vasil'eva, A. Popov // Problems of Infocommunications. Science and Technology : Proc. Internat. Scientific-Practical Conf., Kharkiv, Ukraine, 9-12 Oct. 2018. – Kharkiv, 2018. – P. 281-284. doi: 10.1109/ INFOCOMMST.2018.8632110.

References

1. Fukunaga, K. *Vvedenie v statisticheskuyu teoriyu raspoznavaniya obrazov* [Introduction to statistical theory of pattern recognition]. Moscow, Nauka Publ., 1979. 368 p.

2. Aivazyan, S. A., Bukhshtaber, V. M., Enyukov, I. S., Meshalkin, L. D. *Prikladnaya statistika: klassifikatsiya i snizhenie razmernosti* [Applied Statistics: Classification and Dimension Reduction]. Moscow, Finansy i statistika Publ., 1989. 608 p.

3. He, W., Yu, R., Zheng, Y. and Jiang, T. Image Denoising Using Asymmetric Gaussian Mixture Models. *Internat. Symposium in Sensing and Instrumentation in IoT Era (ISSI)*, Shanghai, 2018, pp. 1-4. doi: 10.1109/ISSI.2018.8538279.

4. Sun, J., Zhao, Y., Wang, S. Image Compression Using GMM Model Optimization. *IEEE Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 1797-1801. doi: 10.1109/ICASSP.2019.8683784.

5. Yang, G., Li, H., Yang, W., Fu, K., Celik, T. and Emery, W. J. Variational Bayesian Change Detection of Remote Sensing Images Based on Spatially Variant Gaussian Mixture Model and Separability Criterion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, vol. 12, no. 3, pp. 849-861. doi: 10.1109/JSTARS.2019.2896233.

6. Bouhlel, N., Méric, S. Unsupervised Segmentation of Multilook Polarimetric Synthetic Aperture Radar Images. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, pp. 1-15. doi: 10.1109/TGRS. 2019.2904401.

7. Shahin, I., Nassif, A. B. and Hamsa, S. Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network. *IEEE Access*, 2019, vol. 7, pp. 26777-26787. doi: 10.1109/ACCESS.2019. 2901352.

8. Hollaus, F., Diem, M. and Sablatnig, R. Multi-Spectral Image Binarization using GMMs. *16th Internat. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, Niagara Falls, NY, 2018, pp. 570-575. doi: 10.1109/ICFHR-2018.2018.00105.

9. Popov, A. V., Pogrebnyak, O., Brashevan, A. N. Image clustering algorithm using polynormal distribution. *Proc. of SPIE Mathematical Methods in Pattern and Image Analysis*, 2005, vol. 5916, pp. 1-9.

10. Vasil'eva, I. K. Iteratsionnyi metod otsenki parametrov smesi funktsii Gaussa v zadachakh opisaniya dannykh nablyudenii [An iterative method for estimating the parameters of a mixture of Gauss functions in problems of describing observational data]. *Radioelektronni i komp'juterni systemy – Radioelectronic and computer systems*, 2015, no. 3 (73), pp. 70-76. (In Russian).

11. Vasil'eva, I., Popov, A. Multicomponent Model of Objects Attributive Signatures on Color Pictures. *Proc. Internat. Scientific-Practical Conf. on Problems of Infocommunications. Science and Technology*, Kharkiv, Ukraine, 9-12 Oct. 2018, pp. 281-284. doi: 10.1109/INFOCOMMST.2018.8632110.

Поступила в редакцию 16.05.2019, рассмотрена на редколлегии 12.06.2019

МЕТОД АВТОМАТИЧНОЇ КЛАСТЕРИЗАЦІЇ ДАНИХ ДИСТАНЦІЙНОГО ЗОНДУВАННЯ

І. К. Васильєва, А. В. Попов

Предметом вивчення в статті є методи автоматичної кластеризації даних дистанційного зондування в умовах апріорної невизначеності щодо кількості класів спостережуваних об'єктів і статистичних характеристик ознак класів. **Метою** є розробка методу апроксимації багатомодальних емпіричних розподілів даних спостережень для побудови вирішальних правил в процедурах попіксельної статистичної класифікації, а також дослідження ефективності даного методу щодо автоматичної класифікації об'єктів на синтезованих і реальних зображеннях. **Завдання**: розробити і реалізувати процедуру розщеплення суміші базисних розподілів, забезпечивши при цьому наступні вимоги: відсутність етапу попереднього аналізу даних з метою вибору оптимальних початкових наближень; добру збіжність методу та можливість автоматичного уточнення переліку класів шляхом об'єднання в один кластер нерозрізнюваних або слабко розрізнюваних компонент суміші; синтезувати тестові зображення із заданою кількістю об'єктів і відомими законами розподілу даних для кожного об'єкта; оцінити ефективність розробленого методу автоматичної класифікації за критерієм імовірності правильного розпізнавання; оцінити результати автоматичної кластеризації реальних зображень.

Використовуваними **методами** є: методи стохастичного моделювання, методи апроксимації емпіричних розподілів, статистичні методи розпізнавання, методи теорії ймовірностей і математичної статистики. Отримані наступні **результати**. Запропоновано метод автоматичного розщеплення суміші розподілів Гауса для побудови порогів прийняття рішення за критерієм максимуму апостеріорної ймовірності. Наведено результати автоматичного формування переліку класів та їх імовірнісних описів, а також результати кластеризації тестових зображень і супутникового знімка. Показано, що розроблений метод є досить ефективним і може застосовуватися для визначення кількості класів об'єктів і математичного опису їх стохастичних характеристик в задачах розпізнавання образів і кластерного аналізу. **Висновки**. Наукова новизна отриманих результатів полягає в тому, що запропонований підхід дозволяє безпосередньо в рамках процедури навчання «без вчителя» оцінювати розрізненість класів і виключати об'єкти, які не розрізняються, із переліку класів.

Ключові слова: розпізнавання образів; кластеризація; апроксимація; суміш базисних функцій; оцінка параметрів суміші; ймовірність правильного розпізнавання.

METHOD FOR AUTOMATIC CLUSTERING OF REMOTE SENSING DATA

I. K. Vasilyeva, A. V. Popov

The subject matter of the article is the methods of automatic clustering of remote sensing data under conditions of a priori uncertainty regarding the number of observed object classes and the statistical characteristics of the signatures of classes. The aim is to develop a method for approximating multimodal empirical distributions of observational data to construct decision rules for pixel-by-pixel statistical classification procedures, as well as to investigate the effectiveness of this method for automatically classifying objects on synthesized and real images. The tasks to be solved are: to develop and implement a procedure for splitting a mixture of basic distributions, while ensuring the following requirements: the absence of a preliminary data analysis stage in order to select optimal initial approximations; a good convergence of the method and the ability to automatically refine the list of classes by combining indistinguishable or poorly distinguishable components of the mixture into a single cluster; to synthesize test images with a specified number of objects and known data distributions for each object; to evaluate the effectiveness of the developed method for automatic classification by the criterion of the probability of correct recognition; to evaluate the results of automatic clustering of real images. The methods used are methods of stochastic simulation, methods of approximation of empirical distributions, statistical methods of recognition, methods of probability theory and mathematical statistics. The following results have been obtained. A method for automatic splitting of a mixture of Gaussian distributions to construct decision thresholds according to the maximal a posteriori probability criterion was proposed. The results of the automatic forming the list of classes and their probabilistic descriptions, as well as the results of the clustering both test images and satellite ones are given. It is shown that the developed method is quite effective and can be used to determine the number of objects' classes as well as their stochastic characteristics' mathematical description for pattern recognition tasks and cluster analysis. Conclusions. The scientific novelty of the results obtained is that the proposed approach makes it possible directly during the "unsupervised" training procedure to evaluate the distinguishability of classes and exclude indistinguishable objects from the list of classes.

Keywords: pattern recognition; clustering; approximation; a mixture of basis functions; estimation of the mixture parameters; the probability of correct recognition.

Васильева Ирина Карловна – канд. техн. наук, доцент, доцент кафедры радиоэлектронных и биомедицинских компьютеризированных средств и технологий, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.

Попов Анатолий Владиславович – канд. техн. наук, доцент, доцент кафедры радиоэлектронных и биомедицинских компьютеризированных средств и технологий, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.

Vasilieva Irina – cand. tehn. sciences, docent, associate professor of the Department of Radio-Electronic and Biomedical Computerized Means and Technologies, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: i.vasilieva@khai.edu, ORCID Author ID: 0000-0002-1378-1104.

Popov Anatolii – cand. tehn. sciences, docent, associate professor of the Department of Radio-Electronic and Biomedical Computerized Means and Technologies, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: a.v.popov@khai.edu, ORCID Author ID: 0000-0003-0715-3870.