УДК 621.396.96

**А.В. ПОПОВ, А.Н. БРАШЕВАН**

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

# АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ НА ОСНОВЕ ТЕОРЕТИКО-ИНФОРМАЦИОННЫХ КРИТЕРИЕВ

*Предложен метод автоматической классификации данных дистанционного зондирования в условиях неполной априорной информации, основанный на теоретико-информационных критериях минимально-достаточного количества информации. Разработан алгоритм кластеризации данных, обеспечивающий устойчивую классификацию всех предъявленных объектов на тестовом изображении с известным количеством объектов и известными их параметрами. Проверка эффективности алгоритма на реальных данных дистанционного зондирования показала, что количество выделяемых кластеров существенно зависит от качества исходного изображения.*

***Ключевые слова:*** *дистанционное зондирование, классификация объектов, автоматическая кластеризация, теория информации.*

## Введение

Системы дистанционного зондирования (ДЗ) с аэрокосмических носителей широко применяются сегодня при решении задач экологического мониторинга окружающей среды, картографирования и предупреждения чрезвычайных ситуаций [1 – 3].

Анализ аэрокосмических снимков позволяет обнаруживать загрязнения водной поверхности, например, разливы нефти, контролировать границы рек, водохранилищ, обнаруживать ледяные заторы, оценивать запасы влаги, состояние и степень эрозии почв и т.д. [1, 2]. Аэрокосмическая съемка поверхности Земли производится как радиолокационными средствами, в т.ч. на различных поляризациях зондирования, так в оптическом и инфракрасном диапазонах [2].

Данные ДЗ представляются, как правило, в виде многоканальных изображений, на которых либо градациями яркости отображается амплитуда сигнала, отраженного от зондируемой поверхности [1, 3], либо применяется цветовое кодирование признаков объектов ДЗ [3, 4].

Большой поток информации обуславливает необходимость автоматической обработки данных ДЗ, одной из задач которой является классификация объектов, информация о которых содержится в получаемых космических снимках. Одним из путей автоматизации этого процесса является кластеризация, то есть, отнесение объектов ДЗ к определённым классам. Автоматическая кластеризация изображения позволяет определить количество классов объектов, различимых на анализируемом изображении. При этом могут быть определены статистические характеристики различимых объектов для последующего их распознавания как на анализируемом, так и на аналогичных изображениях [4, 5].

## Постановка задачи исследований

Известен ряд методов кластеризации [6,7], однако основным их недостатком является недостаточная обоснованность определения количества классов объектов [8]. Автоматическая кластеризация может быть реализована за счет анализа локальных гистограмм, вычисление которых выполняется для скользящего по изображению окна (рис. 1), размер окна при этом должен быть согласован с разрешающей способностью средств дистанционного зондирования [9]. Поскольку на выделенном фрагменте изображения может присутствовать несколько объектов, гистограмма может быть многомодовой (рис. 2). Анализ модового состава гистограммы позволяет определить количество объектов, попавших в выделенный фрагмент изображения, однако при этом всегда существует вероятность пропуска малоразмерных объектов [10].

В данной работе на основе математически строгих критериев минимально достаточного количества информации [11] и минимально достаточной дивергенции [12] предлагается метод определения количества классов объектов, различимых в анализируемых данных ДЗ, а также оценки параметров их статистических моделей.

Целью данной работы является разработка метода автоматической классификации данных ДЗ, не требующего априорной информации о количестве и свойствах наблюдаемых объектов.

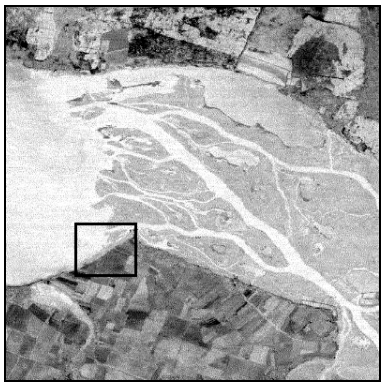


Рис. 1. Изображение данных ДЗ с выделенным на нём фрагментом

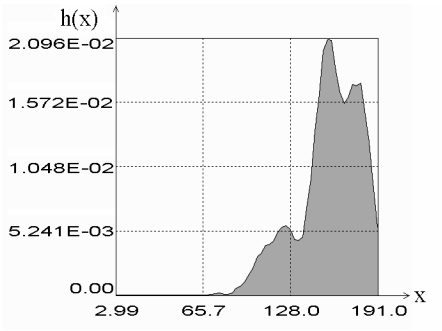



Рис 2. Гистограмма фрагмента изображения, выделенного на рис. 1

## 1. Теоретико-информационное описание объектов ДЗ

Предположим, в данных ДЗ имеется множество классов объектов $A=\{a_k\}$, $k=\overline{1,K}$, которое характеризуется дискретным распределением априорных вероятностей $P=P(a_k)$, $k=\overline{1,K}$, $\sum_{k=1}^{K} P(a_k)=1$. Информация о каждом классе объектов содержится в параметрах $\vec{\xi}=\{\xi_1,...,\xi_m\}$ регистрируемого сигнала $S(t,\vec{\xi})$. Каждый класс $a_k$ описывается плотностью распределения $f(\vec{x}\mid a_k)$, где вектором $\vec{x}$ обозначена конкретная реализация случайного вектора $\vec{\xi}$. Система распознавания объектов ДЗ по результатам измерений параметров $\vec{x}$ сигнала $S(t,\vec{\xi})$ должна определить класс $a_k$.

С точки зрения системы ДЗ наблюдается один случайный процесс $S(t,\vec{\xi})$, являющийся функцией двух случайных величин:

– наблюдение класса объектов $a_k \in \{A\}$ с вероятностной мерой $P=P(a_k)$;

– значение параметров $\vec{x}$ сигнала $S(t,\vec{\xi})$, зависящее от $a_k$, но являющееся случайной величиной при фиксированном $a_k$ в силу флуктуации параметров объектов ДЗ.

Совместное вероятностное пространство $\{\vec{\xi} \cup A\}$ может быть описано дискретно - непрерывным распределением вероятностей $f(A,\vec{x})$, которым описываются любые возможные значения $a_k$ и $\vec{x}$ выборочного пространства $\{A,\vec{\xi}\}$ [12].

При этом маргинальная плотность распределения

$$f(\vec{x})=\sum_{k=1}^{K} P(a_k)\cdot f(A,\vec{x}). \qquad (1)$$

Если $P(a_k)>0$, $k=\overline{1,K}$, то условная вероятность того, что исходом $A$ является $a_k$, при условии, что исходом $\vec{\xi}$ является $\vec{x}$ определяется [6] как

$$P(a_k\mid\vec{x})=\frac{P(a_k)\cdot f(\vec{x}\mid a_k)}{f(\vec{x})}. \qquad (2)$$

Таким образом, появление события $\vec{\xi}=\vec{x}$ изменяет вероятность события $A=a_k$ от априорной вероятности $P(a_k)$ до апостериорной $P(a_k\mid\vec{x})$. Количественной мерой этого изменения принимают [11] неопределенность $H(\bullet)$, которую удается снять при "трансформации" априорного распределения в апостериорное.

$$I(a_k;\vec{x})=H[P(a_k)]-H[P(a_k\mid\vec{x})]=I(a_k)-I(a_k\mid\vec{x}). \qquad (3)$$

Выражение (3) можно интерпретировать как информацию о событии $A=a_k$, содержащуюся в событии $\vec{\xi}=\vec{x}$ [3]. В частности, для меры неопределённости Шеннона [11]

$$I(a_k;\vec{x})=\ln\frac{f(\vec{x}\mid a_k)}{f(\vec{x})}. \qquad (4)$$

Очевидно, что взаимная информация (4) является случайной величиной, т.е. числовой функцией элементов выборочного пространства, но ее значения зависят также от вероятностной меры.

Практически интересным частным случаем взаимной информации (4) является $I(a_k;\vec{x})$ при $P(a_k\mid\vec{x})=1$. При этом $I(A;\vec{x})=H[P(a_k)]$ представляет собой собственную информацию, требуемую для определения $A=a_k$, которая для меры Шеннона равна

$$I(a_k)=\ln\frac{1-P(a_k)}{P(a_k)}=\psi_k. \qquad (5)$$

Собственная информация (5) может быть интерпретирована как априорная неопределенность события $A=a_k$, либо как информация, требуемая для разрешения данной неопределенности. Поэтому в работе [11] $I(a_k)$ (5) обозначено как $\psi_k$ и опре-

делено как минимально достаточное количество информации (МДКИ).

Количество информации, содержащееся в событии $A = a_k$ при условии появления события $\vec{\xi} = \vec{x}$ для меры информации Шеннона равно [11]

$$I(a_k \mid \vec{x}) = \ln\left(\frac{1}{P(a_k) + Q(\vec{x}\,|\,a_k)}\right), \qquad (6)$$

где

$$Q(\vec{x}\,|\,a_k) = \frac{f(\vec{x}\,|\,a_k)}{\sum\limits_{j\neq k=1}^{K} P(a_j)\cdot f(\vec{x}\,|\,a_j)}.$$

Условная собственная информация (6) может быть интерпретирована как информация, которую нужно сообщить наблюдателю для определения $a_k$ после того, как он установил, что произошло событие $\vec{\xi} = \vec{x}$. Очевидно, что (6) эквивалентно апостериорной неопределенности.

## 2. Различимость классов объектов

Традиционным критерием различимости классов объектов на множестве признаков $\vec{\xi}$ является вероятность правильного распознавания множества классов $\{A\}$ [6], для определения которой необходимо выполнить построение решающих правил, что требует знания всех плотностей распределения $f(\vec{x} \mid a_k)$, $k=\overline{1,K}$. В условиях априорной неопределенности относительно количества классов для определения степени их различия необходимо использование методов, позволяющих оценивать различимость классов объектов до построения решающих правил. Для этого предлагается использовать дивергенцию Кульбака [13], которая определяется как мера различимости классов объектов $a_k$ и $a_j$:

$$J(k, j) = \int_{-\infty}^{\infty} \left\{ f\left(\vec{x}\,|\,a_k\right) - f\left(\vec{x}\,|\,a_j\right) \right\} \cdot \ln\left[\frac{P(a_k\,|\,\vec{x})}{P(a_j\,|\,\vec{x})}\right] d\vec{x}. \quad (7)$$

Дивергенция Кульбака (7) является удобной мерой информативности множества параметров сигнала, поскольку обладает следующими свойствами [6]:

1) $J(i, j) > 0$ при $i \neq j$;

2) $J(i, j) = 0$ при $i = j$;

3) $J(i, j) = J(j, i)$;

4) при независимых измерениях $x_1, ..., x_m$ дивергенция аддитивна:

$$J\left(i, j; x_1, x_2, .., x_m\right) = \sum_{k=1}^{m} J(i, j; x_k);$$

5) добавление результата нового измерения $x_{m+1}$ не приводит к уменьшению дивергенции:

$$J\left(i, j; x_1, x_2, .., x_m\right) \le J\left(i, j; x_1, x_2, .., x_m, x_{m+1}\right).$$

Указанные свойства позволяют использовать дивергенцию Кульбака для оценки информативности параметров сигналов, однако (7) применимо лишь в случае двух классов объектов. Для случая $K > 2$ классов в работе [12] вводится понятие обобщенной дивергенции, которая может использоваться для сравнения информативности совокупности параметров сигнала $\{\vec{\xi}\}$ при количестве классов объектов $K > 2$:

$$J\{\vec{\xi}\} = \sum_{k=1}^{K} P(a_k)\cdot J(k\,|\,\vec{\xi}), \qquad (8)$$

где $J(k\,|\,\vec{\xi})$ – мера различимости класса $a_k$ на фоне остальных классов на множестве признаков $\vec{x} \in \vec{\xi}$

$$J(k\,|\,\vec{\xi}) = \int_{-\infty}^{\infty} \left\{ f\left(\vec{x}\,|\,a_k\right) - f\left(\vec{x}\,|\,\breve{a}_k\right) \right\} \cdot \ln\frac{f\left(\vec{x}\,|\,a_k\right)}{f\left(\vec{x}\,|\,\breve{a}_k\right)}\, d\vec{x}. \quad (9)$$

Выражение (9) получено в [12] из (7) путем введения понятия смеси $f(\vec{x}\,|\,\breve{a}_k)$ плотностей вероятностей всех классов, за исключением $a_k$,

$$f(\vec{x}\,|\,\breve{a}_k) = \frac{1}{1 - P(a_k)} \sum_{i\neq k=1}^{K} P(a_i)\cdot f(\vec{x}\mid a_i). \quad (10)$$

Таким образом, соотношения (8) – (10) дают возможность оценить различимость классов объектов при их количестве $K > 2$.

## 3. Минимально достаточная различимость

Как было показано в п.1, в [11] определена связь информативности (4) с МДКИ (5). Совокупность параметров сигнала $\{\vec{\xi}\}$ признается достаточно информативной для распознавания класса $a_k$, если имеются значения $\vec{\xi} = \vec{x}$, для которых выполняется условие

$$I(a_k \mid \vec{x}) \ge \psi_k. \qquad (11)$$

Совокупность параметров сигнала $\{\vec{\xi}\}$ является неинформативной для распознавания класса объектов $a_k$, если при любых значениях $\vec{\xi} = \vec{x}$

$$I(a_k \mid \vec{x}) \le 0.$$

В работе [12] рассмотрена взаимосвязь МДКИ (5) с дивергенцией (8), определено понятие минимально достаточной дивергенции (МДД) в виде

$$\Theta_k = \left[\frac{(1 - P(a_k))^2}{P(a_k)^2} - 1\right] \cdot \ln\left[\frac{(1 - P(a_k))^2}{P(a_k)^2}\right], \quad (12)$$

и доказано, что совокупность параметров сигнала $\{\vec{\xi}\}$ является достаточно информативной в смысле критерия (11) для распознавания класса $a_k$ на фоне остальных классов, если

$$J(k|\vec{\xi}) \geq \Theta_k , \qquad (13)$$

и недостаточно информативной, если условие (13) не выполняется.

Критерий МДД (13) позволяет оценить степень различимости класса объектов $a_k$ на фоне остальных классов, однако требует наличия полной априорной информации.

## 4. Методология классификации в условиях априорной неопределенности

Проблема классификации данных ДЗ связана с решением следующих задач:

– определение количества классов объектов, различимых в анализируемых данных;

– определение параметров статистической модели для каждого из классов;

– построение решающих правил, обеспечивающих классификацию заданного множества объектов.

Классические методы распознавания основаны на отношении правдоподобия [2,6]

$$\frac{f(\vec{x}|a_2)}{f(\vec{x}|a_1)} \underset{<}{>} \frac{P(a_1)}{P(a_2)},$$

из чего следует, что для принятия решения необходимо знать законы распределения $f(\vec{x}\,|\,a_k)$ и априорные вероятности $P(a_k)$ для всех возможных классов объектов $a_k$, $k=\overline{1,K}$, т.е. должно иметь место полное множество событий, для которого

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

В случае обработки реальных данных как правило отсутствует достоверная информация о количестве $K$ наблюдаемых классов объектов $a_k$, нет информации об априорных вероятностях их появления $P(a_k)$, а также отсутствует априорная информация о виде и параметрах законов распределения признаков объектов $f(\vec{x}\,|\,a_k)$.

Одним из путей решения проблемы является кластеризация, то есть автоматическое отнесение данных ДЗ к определённым классам объектов согласно некоторым численным критериям. При этом количество выявленных классов объектов $M$ не обязательно будет совпадать с истинным количеством объектов $K$.

Для описания результатов кластеризации одноканальных данных предлагается использовать статистическую модель в виде смеси нормальных распределений [10]

$$f(x) = \sum_{k=1}^{M} p_k \cdot \phi_k(x) = \sum_{k=1}^{M} p_k \frac{\exp\left\{-\dfrac{(x-m_k)^2}{2\sigma_k^2}\right\}}{\sqrt{2\pi\sigma_k^2}}, \qquad (14)$$

где $M$ – количество нормальных ядер $\phi_k(x)$; $m_k$, $\sigma_k$ – параметры k-го нормального распределения $\phi_k(x)$; $p_k$ – весовые коэффициенты, обеспечивающие выполнение требования

$$\int f(x)dx = 1 . \qquad (15)$$

Все параметры модели изначально неизвестны.

Поскольку данные ДЗ, поступающие с аэрокосмических носителей, представляются, как правило, в виде изображений (радиолокационных или оптических), на первом этапе кластеризации необходимо определить максимальное количество классов, различимых на анализируемом изображении. Для этого выполняется сканирование изображения окном, размеры которого согласованы с разрешающей способностью системы ДЗ. Предполагается, что в выделенном фрагменте изображения (см. рис. 1) содержится несколько классов объектов, поэтому в качестве статистической модели плотности распределения данных во фрагменте изображения принимается смесь нормальных распределений вида (14), где $M$ – количество объектов, попавших в окно, $m_k$, $\sigma_k$, $p_k$ – параметры k-го класса объектов, $k=\overline{1,M}$.

Поскольку объекты на изображении имеют ограниченные размеры в пространстве, то при сдвиге окна относительно его первоначального положения в него попадут другие объекты, и в статистической модели (14) изменятся все параметры. Такое предположение даёт основание проводить последовательный информационный анализ, используя меру количества информации Шеннона (4).

Для принятия решения о появлении в анализируемых данных нового класса объектов используем критерий минимально достаточного количества информации (11). Для меры количества информации Шеннона МДКИ будет иметь вид (5).

Появление события $\xi = x_1$ изменяет вероятность события $A = a_k$ от неизвестной нам истинной априорной вероятности $P(a_k)$ до неизвестной апостериорной $P(a_k|x_1)$ (2). Количественной мерой этого изменения можно принять изменение неопределенности $H$, которое наблюдается при "трансформации" априорного распределения в апостериорное.

Если изначально $P(a_k) \to 0$ и $H[P(a_k)] \to \infty$, то согласно (3) и (6) взаимная информация

$I(a_k|x_1)$, содержащаяся в событии $\xi = x_1$, уменьшает исходную неопределенность

$$H[P(a_k)] - H[P(a_k|x_1)] = \infty - I(a_k|x_1). \quad (16)$$

Последующее измерение $\xi = x_2$ уменьшает исходную неопределенность на величину $I(a_k|x_2)$, обеспечивая снижение уровня неопределенности на величину

$$I(a_k; x_1, x_2) = I(a_k|x_1) + I(a_k|x_2),$$

и дает изменение поступающего количества информации о множестве классов объектов $\{A\}$

$$\Delta I(A; x_1, x_2) = I(A|x_1) - I(A|x_2). \quad (17)$$

Нельзя оценить степень снижения априорной неопределенности (16), поскольку априорные вероятности для классов объектов неизвестны и исходная неопределенность $H[P(a_k)] \to \infty$, но можно оценить изменение получаемого количества информации (17).

Поскольку возможность появления любого класса объектов в анализируемом окне предполагается равновероятной, в качестве априорных вероятностей $P(a_k)$ для вычисления порога принятия решения примем значение $P = 1/N$, где N – количество последовательно сравниваемых окон. В этом случае МДКИ (5) принимает вид [11]

$$\psi = \ln(N-1). \quad (18)$$

Если изменение количества информации превышает порог (18)

$$\Delta I(A; \vec{x}_1, \vec{x}_2) > \psi, \quad (19)$$

то может быть принято решение о наличии класса $a_1$ в множестве $\{A\}$. Последующие измерения позволят расширять множество классов объектов $\{A\}$.

## 5. Алгоритм оценки количества классов объектов и их параметров

Вместо дискретно-непрерывного распределения $f(A, \vec{x})$, знание которого необходимо для определения $I(A; \vec{x})$, в (17) будем использовать его маргинальную плотность (1), в качестве оценки которой может использоваться гистограмма данных, попадающих в локальное окно (рис. 1, 2).

В начале последовательного информационного анализа данные, находящиеся в первом анализируемом фрагменте изображения, можно условно считать принадлежащими первому классу объектов. Для двух сравниваемых окон МДКИ (18) будет иметь нулевое значение, поэтому, если сравнивать только 2 окна, то любые, даже незначительные отличия, которые можно считать отклонениями в пределах одного класса, будут согласно (19) восприняты как появление нового класса. Потому для принятия решения о появлении нового класса объектов необходимо использовать как минимум три последовательных фрагмента данных. При $N = 3$ значение порога (18) увеличится и результат можно считать более достоверным. При использовании более 3-х окон значение МДКИ (18) значительно возрастает, что приводит к повышению вероятности пропуска малоразмерных объектов, поскольку количества отсчётов от них будет не достаточно для превышения порога МДКИ.

Если полученное значение количества информации $I(A;\xi)$ на некотором интервале $\xi \in [x_1, x_2]$ (рис. 3) превышает значение $\psi$, принимается решение о появлении в анализируемом фрагменте данных нового класса объектов. Общее количество объектов становится $M = M + 1$. При этом точку превышения порога $x_1$ и точку $x_2$, в которой значения информативности снова становятся меньше порога можно приближенно считать значениями границ объекта на оси $x$ (рис. 4), что даёт возможность оценить параметры k-го ядра статистической модели (14) объекта, где $k = M$ – номер нового объекта с параметрами статистической модели $m_k$, $\sigma_k$.

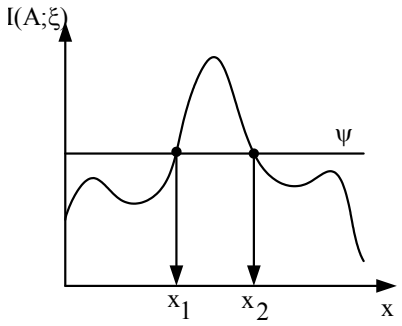



Рис. 3. Зависимость количества информации $I(A, \xi)$ от результатов измерения $\xi = x$ в случае превышения порога $\psi_k$

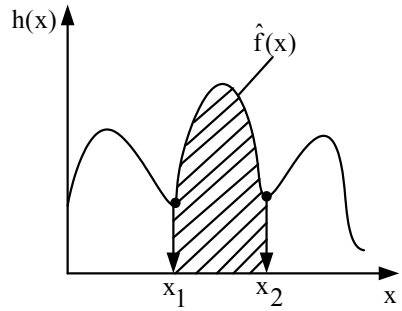



Рис. 4. Интервальная оценка параметров плотности распределения $\hat{f}(x)$ класса объектов по гистограмме $h(x)$

Параметры модели $\hat{f}(x)$ могут быть оценены как [9]

$$\hat{m}_k = \int_{x_1}^{x_2} x \cdot \hat{f}(x)dx ,$$

$$\hat{\sigma}^2 = \int_{x_1}^{x_2} (x - \hat{m}_k)^2 \cdot \hat{f}(x)dx , \qquad (20)$$

где $\hat{f}(x)$ – оценка плотности распределения параметра $x$ на интервале $[x_1, x_2]$, определяемая по фрагменту гистограммы $h(x)$ как

$$\hat{f}(x) = \frac{h(x)}{\Delta x \cdot \sum_{x_1}^{x_2} h(x)}, \qquad (21)$$

где $\Delta x$ – ширина интервала при построении гистограммы $h(x)$.

Последовательный анализ всего изображения по алгоритму (17) – (21) позволяет выделить все объекты на данном изображении, различимые по критерию МДКИ (11), независимо от того, попали они в одно или в разные окна.

## 6. Методика объединения кластеров

Поскольку, как было показано выше, последовательный информационный анализ проводится для трёх соседних фрагментов данных, в полученном списке объектов могут присутствовать повторяющиеся кластеры. На следующем этапе необходимо уточнить количество объектов и исключить повторяющиеся. Для их обнаружения предлагается использовать критерий минимально достаточной различимости. Свойства дивергенции Кульбака (7) позволяют использовать её для оценки степени различимости классов объектов даже при неизвестных априорных вероятностях их появления [8]. Для каждой пары $i, k$ найденных $M$ кластеров может быть определена дивергенция

$$J(i,k) = \int_{-\infty}^{\infty} \left[\hat{f}(x|a_k) - \hat{f}(x|a_i)\right] \cdot \ln \frac{\hat{f}(x|a_k)}{\hat{f}(x|a_i)} \, dx , \qquad (22)$$

значения которой при $i = \overline{1,M}$, $k = \overline{1,M}$ образуют матрицу различимости **J** размером $M \times M$, структура которой представлена на рис. 5. Поскольку из свойств дивергенции следует, что $J(i,j) = J(j,i)$, и $J(i,i) = 0$, то для дальнейших расчётов достаточно использовать только верхний треугольник матрицы **J**.

Для обнаружения слабо различимых классов предлагается использовать критерий минимально достаточной дивергенции для класса $a_k$ на фоне остальных классов (13).

| Номер кластера | 1 | 2 | 3 | ... | M |
|---|---|---|---|---|---|
| 1 | J(1,1) | J(1,2) | J(1,3) | ... | J(1,M) |
| 2 | J(2,1) | J(2,2) | J(2,3) | ... | J(2,M) |
| 3 | J(3,1) | J(3,2) | J(3,3) | ... | J(3,M) |
| ... | ... | ... | ... | | ... |
| M | J(M,1) | J(M,2) | J(M,3) | ... | J(M,M) |

Рис. 5. Матрица различимости классов объектов

Если в процессе анализа матрицы **J** обнаруживаются кластеры, для которых $J(i,j) < \Theta$, то можно предположить, что различие статистических параметров этих кластеров определяется вариацией значений в пределах одного класса объектов, и, следовательно, такие кластеры могут быть объединены в один класс.

Объединение производится путем нахождения оценок статистических характеристик объединяемых данных согласно (20), (21) на интервале $[x_1, x_2]$, общем для объединяемых кластеров.

Таким образом, по критерию минимально достаточной различимости объектов выделяются и объединяются в группы кластеры, различимость которых меньше установленного порога. Различимость оставшихся (в т.ч. объединенных) классов объектов превышает минимально достаточную, и, следовательно, полученное таким способом множество кластеров определяет количество и параметры классов объектов, которые можно различить на анализируемом изображении.

## 7. Построение статистических моделей классов объектов

Полученные на этапе анализа различимости ядер кластеров оценки количества различимых классов объектов $M$ и параметров $m_k$, $\sigma_k$, $k = \overline{1,M}$, распределений $\phi_k(x)$, необходимые для построения модели (14), позволяют определить весовые коэффициенты $p_k$, $k = \overline{1,M}$ для каждой из составляющих смеси распределения (14).

Интерпретация весовых коэффициентов $p_k$ как вероятностей принадлежности данных к кластерам $\phi_k(x)$ позволяет построить алгоритм определения коэффициентов $p_k$ с использованием метода максимального правдоподобия [2]. Для каждого отсчета данных $x_i$ может быть определена вероятность его принадлежности к кластеру $\phi_k(x)$ по критерию максимального правдоподобия [2, 6]:

$$Q_k(x_i) = \frac{\phi_k(x_i)}{\sum_{j=1}^{M} \phi_j(x_i)}. \qquad (18)$$

В качестве параметров функций $\phi_k(x)$ используются ранее найденные оценки $m_k$, $\sigma_k$. Затем определяется максимальное значение вероятности $Q_k(x_i)$, $k = \overline{1,M}$, номер $k$ которой определяет номер кластера $\phi_k(x)$.

Подсчитав количество $n_k$ отсчетов $x_i$, попавших в каждый кластер, можно определить оценки вероятностей

$$p_k = n_k / N,$$

где $N$ – общее количество отсчетов в анализируемых данных.

Разумеется, такой подход дает приближенную оценку вероятностей $p_k$, поскольку неявно предполагается равновероятная принадлежность отсчета $x_i$ к кластерам $\phi_k(x)$, $k = \overline{1,M}$, однако при этом автоматически обеспечивается требование

$$\sum_{k=1}^{M} p_k = 1.$$

Найденные по предлагаемой методике оценки параметров аппроксимирующего многомодового распределения нуждаются в уточнении, для чего следует применить оптимизационные процедуры [10]. В качестве варьируемых параметров при этом используются оценки параметров кластеров $m_k$, $\sigma_k$, $p_k$, $k = \overline{1,M}$.

## 8. Результаты тестирования алгоритма кластеризации

Для оценки эффективности работы алгоритма автоматической кластеризации требуется проверка его на тестовых данных. Для этого использовалось синтезированное изображение (рис. 6) с заранее заданным количеством объектов M=5 и известными законами распределения данных для каждого объекта, параметры которых представлены в табл. 1. Гистограмма тестового изображения и положения объектов представлены на рис. 7.

Таблица 1

Параметры классов тестовых объектов

| k | $m_k$ | $\sigma_k$ | $p_k$ |
|---|---|---|---|
| 1 | 175 | 2,5 | 0,24 |
| 2 | 10 | 3 | $3{,}68\times10^{-2}$ |
| 3 | 100 | 3,5 | $2{,}7\times10^{-3}$ |
| 4 | 146 | 4 | 0,719 |
| 5 | 193 | 4,5 | $2{,}9\times10^{-4}$ |

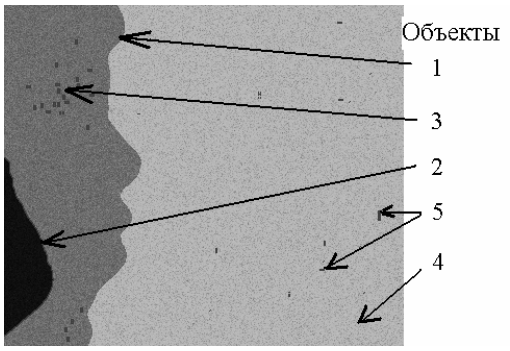



Рис. 6. Тестовое изображение

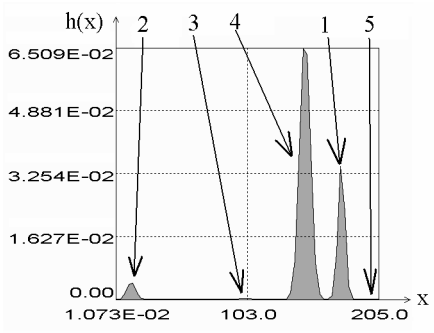



Рис. 7. Гистограмма тестового изображения

В табл. 2 представлены расчетные значения дивергенции Кульбака (7), используемой в качестве меры различимости классов объектов на тестовом изображении.

Таблица 2

Различимость классов тестовых объектов

| Номер кластера | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 133,3 | 133,2 | 86,9 | 48,3 |
| 2 | - | 0 | 132,9 | 132,8 | 132,6 |
| 3 | - | - | 0 | 132,0 | 132,5 |
| 4 | - | - | - | 0 | 124,9 |
| 5 | - | - | - | - | 0 |

Минимально достаточная различимость (12) для рассматриваемого случая ($M = 5$) имеет значение $\Theta = 41{,}6$, расчетные величины дивергенции Кульбака (табл. 2) значительно превышают этот порог, следовательно, все классы объектов на рис. 6 достаточно различимы в смысле критерия (11).

Для определения количества объектов выполнялся последовательный просмотр изображения по фрагментам, попавшим в «скользящее» окно. На рис. 8 показаны гистограммы, построенные для 2-х произвольно выбранных окон.

Последовательный информационный анализ согласно (11) в каждой точке $x \in [x_{min}, x_{max}]$ позволил определить информативность данных, попавших в анализируемое окно, по сравнению с предыдущим окном. Для принятия решения о появле-

нии нового класса по сравнению с предыдущим используется порог минимально достаточного количества информации (18).

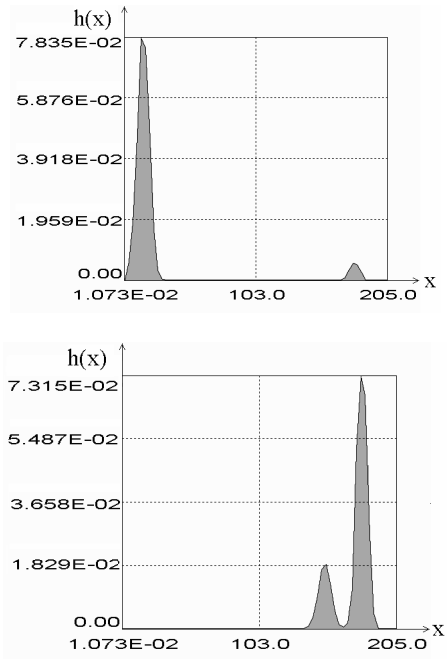
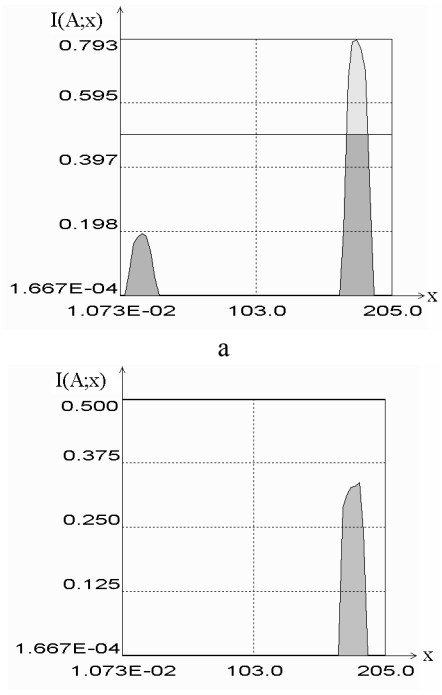



Рис. 8. Гистограммы произвольно выбранных фрагментов изображения (рис.6)

На рис. 9, а показана зависимость информативности I(x) относительно порога МДКИ $\psi$ при появлении нового класса. На рис. 9, б показан случай, когда для принятия решения о появлении нового класса информации не достаточно.




а

б

Рис. 9. Зависимость информативности I(x)

В результате выполнения первой части алгоритма, рассмотренной в п. 5, на рис. 6 было обнаружено 90 кластеров. После объединения (п. 6) и оценки параметров кластеров (п. 7) получено количество классов, равное 5, параметры которых представлены в табл. 3.

Таблица 3

Оценки параметров кластеров

| k | $m_k$ | $\sigma_k$ | $p_k$ |
|---|---|---|---|
| 1 | 179,9 | 3 | 0,27 |
| 2 | 15 | 3,5 | 0,014 |
| 3 | 105 | 3,9 | $6{,}5\times10^{-3}$ |
| 4 | 151 | 4,4 | 0,579 |
| 5 | 198 | 3,5 | $1{,}4\times10^{-3}$ |

Оценки параметров нормальных распределений (14), полученные после разделения выборки на классы и представленные в табл. 3., используются в качестве априорной информации в процедуре принятия решений при распознавании классов объектов на изображении (рис. 6) по критерию максимума апостериорной вероятности (2). Результат распознавания объектов на тестовом изображении представлен на рис. 10.

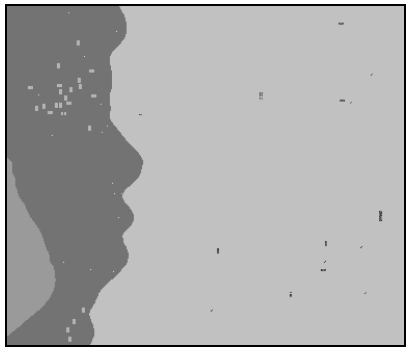


Рис. 10 . Результат кластеризации исходного изображения (рис. 6)

Для оценки достоверности результатов классификации были рассчитана матрица решений. Она представляет собой таблицу соответствия найденных классов заданным. Количество принятых решений в пользу каждого из классов объектов нормировано на количество пикселей в исходном изображении. При этом сумма чисел в диагонали таблицы соответствует вероятности правильного распознавания (0,99953), а сумма остальных значений соответствует вероятности ошибочной классификации ($4{,}7\times10^{-4}$). Необходимо отметить, что вероятности правильных решений в пользу каждого класса (табл. 4) с высокой точностью соответствуют априорным вероятностям классов объектов (табл. 1). Таким образом, результаты тестирования алгоритма автоматической классификации данных подтверждают его высокую эффективность.

Таблица 4

Матрица вероятностей принятых решений

| Предъявлен класс | Приняты решения | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0,24 | 0 | 0 | 0 | $1{,}27\times\times10^{-4}$ |
| 2 | 0 | $3{,}68\times\times10^{-2}$ | 0 | 0 | 0 |
| 3 | 0 | 0 | $2{,}72\times\times10^{-3}$ | $7{,}96\times\times10^{-6}$ | 0 |
| 4 | $3{,}34\times\times10^{-4}$ | 0 | 0 | 0,719 | 0 |
| 5 | 0 | 0 | 0 | 0 | $7{,}96\times\times10^{-3}$ |

## 9. Апробация метода автоматической кластеризации на реальных данных

Предложенный метод кластеризации данных на основе критериев информационной достаточности был применён для обработки реальных данных ДЗ. В качестве примера было взято изображение, представленное на рис. 1. Автоматическим классификатором на данном изображении было обнаружено 12 различимых кластеров, каждый из которых представлен на рис. 11 своей интенсивностью.

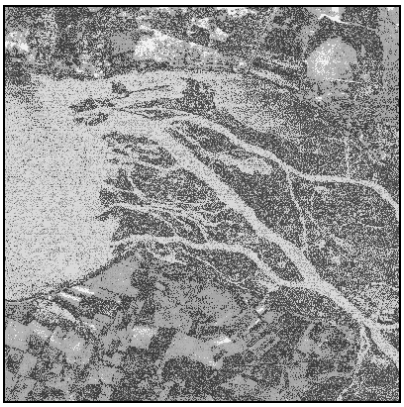


Рис. 11. Результат кластеризации изображения (рис. 1)

Сопоставление рис. 1 и рис. 11 показывает, что автоматическая классификация данных ДЗ обеспечивает отнесение к одному классу объектов участков изображения, содержащих данные об объектах одной физической природы. Следует отметить, что зашумленность реальных данных приводит к выделению дополнительных кластеров, однако соответствие этих кластеров шумовым компонентам или внутриклассовым флуктуациям сигнала требует дальнейших исследований.

## Заключение

Одним из методов решения задачи распознавания объектов на радиолокационных изображениях в условиях априорной неопределенности относительно количества классов объектов и их статистических свойств является автоматическая кластеризация изображений, которая позволяет выделить на изображении различимые классы объектов.

Разработан метод автоматической классификации данных ДЗ в условиях априорной неопределенности, математической основой которого являются критерии достаточного количества информации. Задача разделения данных на отдельные кластера решается за счет использования критерия минимально достаточного количества информации. Уточнение количества реально различимых объектов на изображении и объединение кластеров выполняются на основе критерия минимально достаточной различимости. Поскольку априорная информация о законах распределения параметров сигналов отсутствует, в качестве статистической модели для каждого из найденных классов объектов принимается нормальное распределение, параметры которого служат описанием выделенных кластеров.

Результаты тестирования алгоритма кластеризации на модельных данных показали, что параметры выделенных кластеров могут использоваться в качестве статистической модели при распознавании объектов, содержащихся в данных ДЗ.

Недостатком предложенного алгоритма является использование гипотезы о нормальном распределении параметров объектов, что может не соответствовать реальности.

Применение предложенного алгоритма для кластеризации реальных радиолокационных изображений продемонстрировало его устойчивость. Однако качество исходного изображения существенно влияет на количество выделенных объектов и точность определения их параметров.

## Литература

*1. Радиолокационные методы и средства оперативного дистанционного зондирования Земли с аэрокосмических носителей / Под ред. С.Н. Конюхова, В.И. Драновского, В.Н. Цимбала. – К.: НАНУ, 2007. – 440 с.*

*2. Волосюк В.К. Статистическая теория радиотехнических систем дистанционного зондирования и радиолокации / В.К. Волосюк, В.Ф. Кравченко. – М.: ФИЗМАТЛИТ, 2008. – 704 с.*

*3. Красовский Г.Я. Введение в методы космического мониторинга окружающей среды /Г.Я. Красовский, В.А. Петросов. – Х.: Гос. Аэрокосм. ун-т*

*им. Н.Е. Жуковского «ХАИ», 1999. – 205 с.*

*4. Lee J. Unsupervised Classification Using Polarimetric Decomposition and the Complex Wishart Classifier /Jong-Sen Lee, M.R. Grunes, T. L. Ainsworth, L.J. Du, D. L. Schuler, S. R. Cloude // IEEE Trans. on Geoscience and Remote Sensing. – 1999. – V. 37, N. 5. – P. 2249-2258.*

*5. Robin A. Unsupervised Subpixelic Classification Using Coarse-Resolution Time Series and Structural Information / A. Robin, S. Le Hégarat-Mascle, L. Moisan // IEEE Trans. on Geoscience and Remote Sensing. – 2008. – V. 46, N. 5. – P. 1359-1373.*

*6. Фукунага К. Ведение в статистическую теорию распознавания образов: пер. с англ. / К. Фукунага. – М.: Наука, 1979. – 367 с.*

*7. Kaufman L. Finding Groups in Data. An Introduction to Cluster Analysis. / L. Kaufman, P.J. Rousseeuw. – NY: A Wiley-Interscience Publ., 1989. – 331 p.*

*8. Information Theory-Based Approach for Contrast Analysis in Polarimetric and/or Interferometric SAR Images / J. Morio, P. Réfrégier, F. Goudail, P.C. Dubois-Fernandez, X. Dupuis // IEEE Trans. on Geoscience and Remote Sensing. – 2008. – V. 46, N. 8. – P. 2185-2196.*

*9. Mignotte M. Segmentation by Fusion of Histogram-Based K-Means Clusters in Different Color Spaces / M. Mignotte // IEEE Trans. on Image Processing. – 2008. – V. 17, N. 5. – P. 780-787.*

*10. Popov A.V. Image clustering algorithm using polynormal distribution / A.V. Popov, O.B. Pogrebnyak, A.N. Brashevan // Mathematical Methods in Pattern and Image Analysis: Proc. SPIE. – 2005. – V. 5916. – P. 341-349.*

*11. Косенко Г. Г. Критерии информативности при различении сигналов / Г.Г. Косенко. – М.: Радио и связь, 1982. – 214 с.*

*12. Попов А.В. Критерий информативности параметров сигналов для радиолокационного распознавания объектов / А.В. Попов // Авиационно-космическая техника и технология: сб. научн. трудов Гос. Аэрокосм. ун-та им. Н.Е. Жуковского «ХАИ». – Х., 1999. – Вып. 12. – С. 44-47.*

*13. Кульбак С. Теория информации и статистика: пер. с англ. / С. Кульбак. – М.: Наука, 1968. – 302 с.*

*Поступила в редакцию 10.04.2009*

**Рецензент:** д-р техн. наук, профессор, проф. кафедры производства радиоэлектронных систем Национального аэрокосмического университета им. Н.Е. Жуковского «ХАИ» Г.Я. Красовский, Харьков.

## АВТОМАТИЧНА КЛАСИФІКАЦІЯ ДАНИХ ДИСТАНЦІЙНОГО ЗОНДУВАННЯ НА ОСНОВІ ТЕОРЕТИКО–ІНФОРМАЦІЙНИХ КРИТЕРІЇВ

***А.В. Попов, О.М. Брашеван***

Запропоновано метод автоматичної класифікації даних дистанційного зондування за умов неповної апріорної інформації, що базується на теоретико–інформаційних критеріях мінімально достатньої кількості інформації. Розроблено алгоритм кластеризації даних, що забезпечує стійку класифікацію об'єктів на тестовому зображенні з відомою кількістю об'єктів та відомими їх параметрами. Перевірка ефективності алгоритму на реальних даних дистанційного зондування показала, що кількість виділюваних кластерів істотно залежить від якості вихідного зображення.

**Ключові слова:** дистанційне зондування, класифікація об'єктів, автоматична кластеризація, теорія інформації.

## REMOTE SENSING DATA AUTOMATIC CLASSIFICATION BASED ON THE INFORMATION-THEORETICAL CRITERIA

***A.V. Popov, A.N. Brashevan***

A method for remote sensing data automatic classification in conditions of incomplete a priori information based on the information-theoretical criteria of minimal-sufficient quantity of information is suggested. A data clustering algorithm that provides stable classification of all the presenting objects on the test image with known number of objects and their parameters is developed. Verification of the algorithm's efficiency on real remote sensing data showed that the number of the selected clusters greatly depends on the quality of the initial image.

**Key words:** remote sensing, objects classification, automatic clustering, information theory

**Попов Анатолий Владиславович** – канд. техн. наук, доцент, доцент кафедры производства радиоэлектронных систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина, e-mail: a.v.popov@inbox.ru.

**Брашеван Александра Николаевна** – аспирантка кафедры производства радиоэлектронных систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.