РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2006, том 42, № 2

АНАЛИЗ И СИНТЕЗ СИГНАЛОВ И ИЗОБРАЖЕНИЙ

УДК 004.93'1; 004.932

И. А. Пестунов, Ю. Н. Синявский

(Новосибирск)

НЕПАРАМЕТРИЧЕСКИЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ НА ОСНОВЕ GRID-ПОДХОДА

Предложен быстрый непараметрический алгоритм автоматической классификации для сегментации многоспектральных аэрокосмических изображений. Алгоритм основан на формировании сеточной структуры данных в пространстве спектральных признаков и использовании ее в итеративной процедуре «среднего сдвига» для поиска локальных мод плотности распределения. Представлены результаты тестирования на модельных и реальных аэрокосмических данных.

Введение. При решении задач, связанных с анализом и распознаванием данных дистанционного зондирования (ДДЗ), методы кластеризации занимают одно из центральных мест [1–3]. Алгоритмы кластеризации, реализованные в известных пакетах программ обработки ДДЗ (ERDAS Imagine, ENVI, IDRISI и др.), требуют от пользователя задания ряда параметров, предопределяющих как количество кластеров (классов), так и их форму, размер. На практике пользователи, как правило, не имеют априорной информации, необходимой для выбора этих параметров. Кроме того, простые математические модели, которые лежат в основе этих алгоритмов, не позволяют выделять классы сложной формы, наиболее адекватно отражающие реальные данные. Указанные недостатки часто приводят к неудовлетворительным результатам кластеризации.

Как известно [1], для обработки ДДЗ статистический подход является наиболее адекватным. При этом подходе предполагается, что выборочное пространство данных есть множество реализаций случайной величины, плотность распределения которой неизвестна. Локальные моды этой плотности соответствуют центрам классов, а ее «овраги» определяют границы разделения кластеров. Для оценивания неизвестной плотности распределения целесообразно использовать непараметрические оценки. Преимущество алгоритмов кластеризации, основанных на непараметрических оценках [4, 5], заключается в том, что они не накладывают ограничений на размер и форму выделяемых классов. К недостаткам таких алгоритмов относится их вычислительная сложность (порядка $O(N^2)$, где N – объем выборки).

В работе [6] предложен непараметрический алгоритм, который позволяет за приемлемое время обрабатывать выборки размером в несколько десятков тысяч элементов. Однако в задачах обработки ДДЗ характерный объем выборки – сотни тысяч и даже миллионы элементов.

В данной работе предлагается непараметрический алгоритм кластеризации, основанный на учете особенностей ДДЗ, сеточной структуре данных, формируемой в пространстве спектральных признаков, и процедуре «среднего сдвига», которая порождает естественное разбиение выборки на классы (идея этой процедуры была предложена в работе [7], получила развитие и использована в работах [8–10]).

1. Постановка задачи и метод ее решения. Предположим, что произведена k-зональная съемка участка местности, содержащего N элементов разрешения, тогда результат съемки можно представить в виде множества

$$X = \{x^{(i)} = (x_1^{(i)}, \ldots, x_k^{(i)}) \in R^k,\ \ i = \overline{1, N}\},$$

где $x_j^{(i)}$ – значение яркости i-го элемента разрешения в j-м диапазоне спектра ($j = \overline{1, k}$). Пусть каждый вектор $x^{(i)}$ – реализация k-мерного случайного вектора $\mathbf{x}$, плотность распределения $f(x)$, $x = (x_1, \ldots, x_k) \in R^k$, которого неизвестна, а также нет какой-либо априорной информации о ее параметрическом виде. В этих условиях для оценивания плотности $f(x)$ в точке $x \in R^k$ целесообразно воспользоваться непараметрической парзеновской оценкой $\hat{f}_N(x)$, определяемой выражением

$$\hat{f}_N(x) = \frac{1}{Nh^k} \sum_{i=1}^{N} \Phi\left(\frac{x - x^{(i)}}{h}\right),$$

где h – параметр сглаживания; $\Phi(x)$ – колоколообразная функция (ядро), удовлетворяющая условиям:

$$\Phi(x) \geq 0\ \ \forall x \in R^k, \qquad \sup_{x \in R^k} \Phi(x) < \infty,$$

$$\int_{R^k} \Phi(x)\, dx = 1, \qquad \lim_{\|x\| \to \infty} \|x\|^2 \Phi(x) = 0.$$

Эта оценка простая и, в отличие от гистограммной оценки и оценки k-ближайших соседей, обладает высокими асимптотическими свойствами. Она является несмещенной, состоятельной в среднеквадратическом смысле и равномерно сходящейся по вероятности при условии, что

$$\lim_{N \to \infty} h(N) = 0, \qquad \lim_{N \to \infty} Nh^k(N), \qquad \lim_{N \to \infty} Nh^{2k}(N) = \infty.$$

Пусть

$$m_h(x) = \frac{1}{n_x} \sum_{x^{(i)} \in S_h(x)} x^{(i)}$$

является выборочным средним в точке $x \in R^k$. Здесь $S_h(x)$ – шар с центром в точке x и радиусом h, а n_x – количество точек множества X, содержащихся в $S_h(x)$. Тогда согласно [8, 10] разность $m_h(x) - x$ есть вектор среднего сдвига, который интересен тем, что его направление совпадает с направлением градиента оценки $\hat{f}_N(x)$ в точке x, если в качестве ядра оценки использовать ядро Епанечникова, определяемое выражением

$$\Phi_E(x) = \begin{cases} \frac{1}{2} V_k^{-1}(k+2)(1 - x^T x), & \text{если } x^T x < 1; \\ 0, & \text{если } x^T x \le 1, \end{cases}$$

где V_k – объем единичного k-мерного шара.

Повторяющиеся движения от точки $x \in R^k$ к ее выборочному среднему $m_h(x)$, затем от $x_1 = m_h(x)$ к $m_h(x_1)$ и т. д. до шага n, на котором значение $m_h(x_n)$ будет равно $m_h(x_{n+1})$, называют алгоритмом среднего сдвига. Этот алгоритм представляет собой адаптивную процедуру наискорейшего подъема для нахождения локальных мод плотности $\hat{f}(x)$.

Процедура среднего сдвига порождает естественное разбиение множества X на классы: точки $x^{(i)}$ и $x^{(j)}$ принадлежат к одному классу, если итеративные процессы среднего сдвига, начинающиеся с этих точек, сходятся к одной и той же моде [10]. Такая процедура достаточно трудоемка, поэтому ее непосредственное применение ограничено выборками небольшого объема.

В работе [6] предложен алгоритм кластеризации, в котором процедура среднего сдвига применяется не для всех точек исходного множества X, а лишь для некоторого случайного подмножества из X. Этот прием позволяет увеличить объем обрабатываемых данных до нескольких десятков тысяч.

В типичных задачах тематической обработки аэрокосмических данных число классифицируемых объектов исчисляется сотнями тысяч и миллионами, поэтому использование такого алгоритма для обработки многозональных данных также приводит к неприемлемо большим вычислительным затратам.

В следующем разделе описан быстрый алгоритм кластеризации многозональных данных, в котором стартовое множество точек для запуска процедуры среднего сдвига порождается клеточной структурой данных, формируемой в пространстве спектральных признаков.

2. Описание алгоритма. Предлагаемый алгоритм кластеризации опирается на использование двух характерных особенностей многозональных данных. Первая из них заключается в ограниченности диапазонов изменения значений спектральных признаков (значения лежат в диапазоне целых чисел от 0 до $K-1$, где K – число уровней квантования видеосигнала, обычно не превышающее 256), а вторая – в высокой частоте повторяемости векторов спектральных яркостей. Повторяемость обусловливается ограниченностью диапазона спектральных яркостей, наличием корреляции между спектральными диапазонами, а также относительной однородностью и достаточной протяженностью природных объектов. Алгоритм можно записать в виде последовательности шагов.

1. *Формирование клеточной структуры данных в пространстве спектральных признаков*. Разобьем все пространство значений спектральных признаков $[0,255]_1 \times \ldots \times [0,255]_k$ на гиперкубические клетки со стороной $2h$

(h – параметр сглаживания). Введем общую нумерацию клеток (последовательно от одного слоя клеток к другому) и с каждой клеткой свяжем набор попавших в нее спектральных векторов из X.

2. Формирование таблицы «весов» векторов множества X. Под «весом» вектора **x** понимаем число вхождений x в множество X. Для экономии памяти каждую клетку можно обрабатывать отдельно. Таблица весов позволяет иногда в десятки раз сократить объем вычислений при определении выборочных средних и оценок плотностей распределения.

3. *Формирование множества начальных (стартовых) векторов S для запуска процедуры среднего сдвига.* Для каждой клетки, которая содержит более $N_{\min}$ векторов из X, вычислим вектор средних значений по всем точкам, попавшим в эту клетку. Совокупность полученных таким образом средних векторов образует множество $S = \{s_1, \ldots, s_L\}$.

4. *Оценивание локальных мод плотности $\hat{f}(x)$.* Применим процедуру среднего сдвига, используя в качестве стартовых векторов элементы множества S. При этом стоит заметить, что для вычисления векторов $m_h(x)$ не нужно перебирать все элементы X, а достаточно использовать лишь элементы соседних с точкой x клеток, которые легко определяются благодаря введенной нумерации. В результате применения процедуры получим множество локальных мод $Z_0 = \{z_1, \ldots, z_{M_0}\}$. Каждой точке множества $s \in S$ поставим в соответствие элемент из Z_0, на котором остановилась процедура среднего сдвига, стартовавшая из s.

5. *Распределение точек множества X по классам.* Используя метод ближайшего соседа, а в качестве обучающей выборки размеченное множество S, распределим множество X по классам. Клеточная структура данных обеспечивает организацию быстрого поиска ближайшего соседа.

6. *Формирование множества кандидатов в центры классов* $Z_1 = \{z_1, \ldots, z_{M_1}\}$. Определим на множестве Z_0 все подмножества близких друг к другу точек (точка близка к подмножеству, если она находится на расстоянии не больше h от некоторой точки этого подмножества). Для каждого выделенного подмножества вычислим его вектор средних значений. Совокупность таких векторов и будет образовывать множество Z_1.

7. *Нахождение множества центров классов $Z = \{z_1, \ldots, z_M\}$.* Будем считать, что две точки множества Z_1 связаны, если между ними нет оврага в оценке плотности распределения $\hat{f}_N(x)$. Существование оврага определяется для каждых двух точек z_i, z_j. Вдоль линии, соединяющей точки z_i и z_j (начиная с точки меньшей плотности), с шагом h вычислим оценку плотности $\hat{f}_N(x)$. Если для некоторой точки x_m линии, соединяющей z_i и z_j, выполнено

$$\frac{\max[\hat{f}_N(x_1), \ldots, \hat{f}_N(x_{m-1})]}{\hat{f}_N(x_m)} > T,$$

где

$$\hat{f}_N(x_1) = \min[\hat{f}_N(z_i), \hat{f}_N(z_j)], \quad T \geq 1,$$

то констатируем обнаружение оврага. На заключительном этапе множество Z_1 разбиваем на компоненты связности. Процесс выделения этих компонен-

тов заключается в последовательной обработке всех элементов $z \in Z_1$ процедурой *Process*:

Process(z)
{
Исключаем z из Z_1;
Определяем клетку K, в которой находится z;
Формируем множество $Z_1(K, z)$, состоящее из Z_1, находящихся в соседних с K клетках и связанных с z;
Для всех ($x \in Z_1(K, z)$) выполнить *Process(x)*;
}

В выделенном таким образом компоненте связности найдем элемент с наибольшей плотностью и добавим его к Z. Как показали эксперименты, в оценках $\hat{f}_N(x)$ вместо квадратичного ядра Епанечникова целесообразнее использовать мультипликативное треугольное ядро $\Phi_\Delta(x)$, потому что качество кластеризации при этом не ухудшится, а объем вычислений существенно сократится. Ядро

$$\Phi_\Delta(x) = \prod_{i=1}^{k} \Phi_\Delta^{(1)}(x_i),$$

где

$$\Phi_\Delta^{(1)}(y) = \begin{cases} 1 - |y|, & \text{если } |y| \le 1; \\ 0, & \text{если } |y| > 1, \end{cases} \quad y \in R^1.$$

Так как ядро финитно, при вычислении оценок $\hat{f}_N(x)$, как и при вычислении $m_h(x)$ в п. 4 (разд. 2), достаточно использовать лишь элементы клеток, которые являются соседними с клеткой, содержащей точку x.

Представленный алгоритм имеет вычислительную сложность порядка $O(LN)$.

Заметим, что предложенная схема алгоритма допускает распараллеливание наиболее трудоемких этапов обработки, позволяющее повысить вычислительную эффективность при реализации ее на многопроцессорных вычислительных системах.

3. Результаты экспериментальных исследований. Предложенный алгоритм программно реализован в среде Microsoft Visual Studio .NET и включен в пакет прикладных программ GIPARD [11], который предназначен для автоматизированного анализа ДДЗ. Для работы алгоритма необходимо задать значения трех параметров: h, $N_{\min}$, T. Многочисленные экспериментальные исследования, проведенные как со спутниковыми данными, так и с аэроснимками, показывают, что кластеризацию многозонального изображения целесообразно проводить после применения линейного растяжения динамического диапазона значений яркости в каждой зоне на всю допустимую шкалу. Это преобразование значительно упрощает подбор оптимальных значений для параметра сглаживания h. Эксперименты показали, что оптималь-

ные значения параметра h удовлетворяют неравенству $7 \le h \le 15$. Параметры $N_{\min}$ и T оказывают слабое влияние на результат кластеризации. В экспериментах $N_{\min} = 0$, $1{,}5 \le T \le 2{,}1$. Далее приведены результаты экспериментов на модельных и реальных данных, обработка которых производилась на ПЭВМ "Pentium IV" с тактовой частотой 2,8 ГГц.

Эксперимент 1. Использовались двумерные данные, состоящие из 900 точек, сгруппированных в три линейно неразделимых кластера по 300 точек (рис. 1, *a*). При кластеризации с параметрами $h = 10$, $N_{\min} = 0$, $T = 1{,}95$ выделены четыре класса, содержащие 300, 300, 297 и 3 точки. Результат кластеризации представлен на рис. 1, *b*.

Эксперимент 2. Использовались двумерные данные, состоящие из 1000 точек, сгруппированных в два линейно неразделимых кластера, содержащие 300 и 700 точек (рис. 2, *a*). При кластеризации с параметрами $h = 12{,}5$, $N_{\min} = 0$, $T = 1{,}6$ выделены два класса, содержащие 300 и 700 точек. Результат кластеризации видим на рис. 2, *b*.

Эксперимент 3. Использовались двумерные данные, состоящие из 990 точек, сгруппированных в три нормально распределенных класса, содержащие по 330 точек (рис. 3, *a*) с параметрами $h = 13$, $N_{\min} = 0$, $T = 1{,}7$. Выделено пять классов, содержащих 330, 329, 329, 1 и 1 точку. Результат кластеризации представлен на рис. 3, *b*.

Эксперимент 4. Использовалось цветное изображение (рис. 4, *a*). Кластеризация проводилась в цветовом пространстве $R \times G \times B$. Каждый кластер соответствовал однородной области на изображении. Цветовое пространство содержало 1283800 точек (средний вес точек равнялся 19,69). Выделено 12 классов при параметрах $h = 11$, $N_{\min} = 0$, $T = 1{,}5$. Время обработки составило 11 с. Сегментированное изображение, полученное в результате кластеризации, показано на рис. 4, *b*.

Эксперимент 5. Использовалось изображение Краснотуранского бора (Красноярский край), которое получено с помощью самолетного сканера С-500, разработанного в Институте космических исследований РАН. Сканер представляет собой систему из восьми каналов, которым соответствуют длины волн: 800–970, 620–710, 565–630, 490–550, 525–575, 685–730,

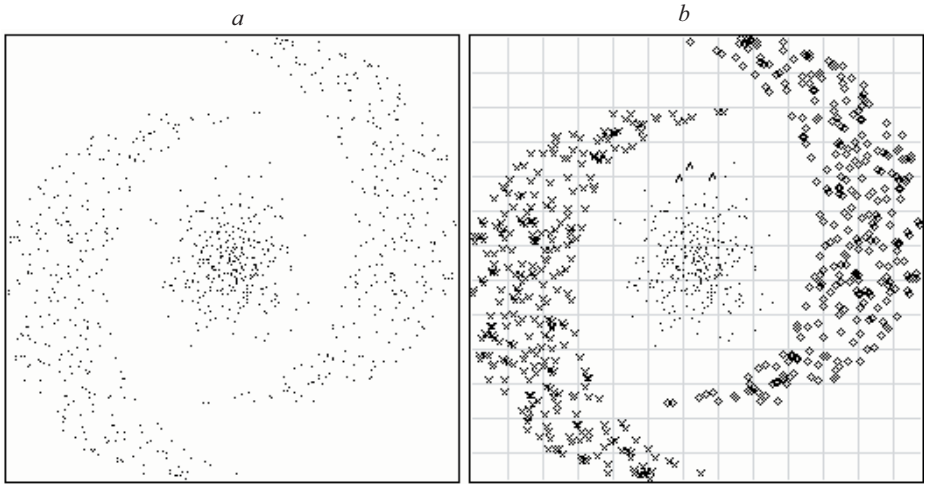

Рис. 1. Статистическое моделирование: исходное множество (*a*), результат кластеризации (*b*)

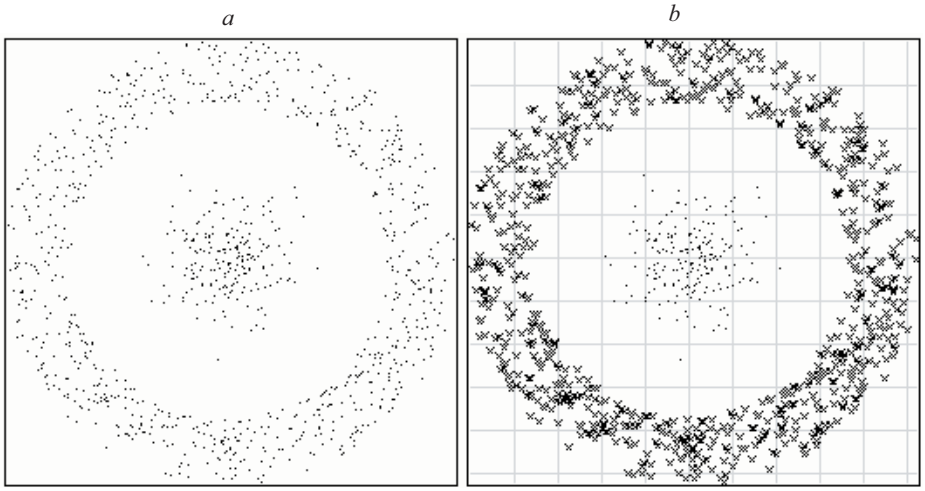

Рис. 2. Статистическое моделирование: исходное множество (*a*), результат кластеризации (*b*)

730–800, 900–1060 нм. Высота съемки 7300 м, размер разрешаемого элемента на местности 9×18 м. Обрабатывался фрагмент размером 256×250 в 1-, 2-, 6- и 7-м каналах. Исходное число элементов 64000, параметры: $h = 13$, $N_{\min} = 0$, $T = 1{,}9$. Время обработки 4 с. Для интерпретации картосхем использовались планы лесонасаждений и результаты визуально-инструментального дешифрирования спектрозональных аэроснимков. Для уменьшения раздробленности картосхем, получаемых в ходе попиксельной классификации многозональных изображений, целесообразно производить их постобработку медианным фильтром. Качество картосхем оценивалось специалистами-дешифровщиками и было признано удовлетворительным. Результаты эксперимента представлены на рис. 5.

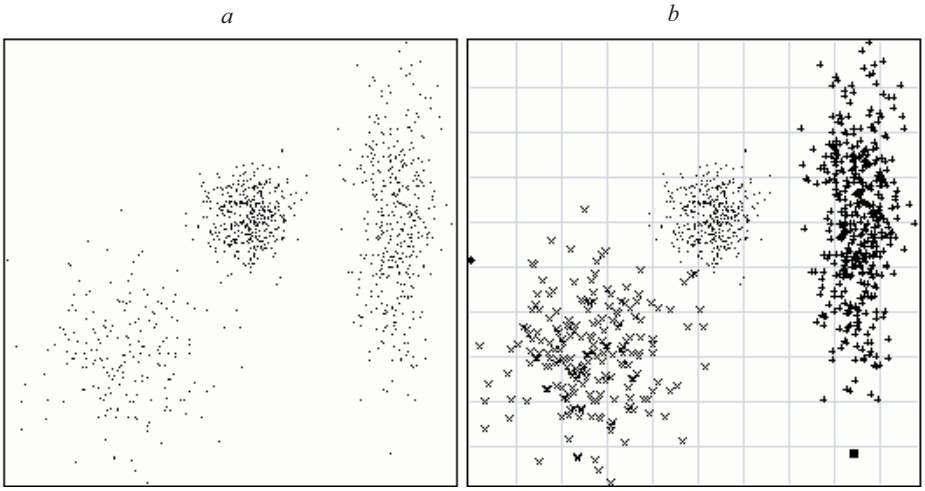

Рис. 3. Статистическое моделирование: исходное множество (*a*), результат кластеризации (*b*)

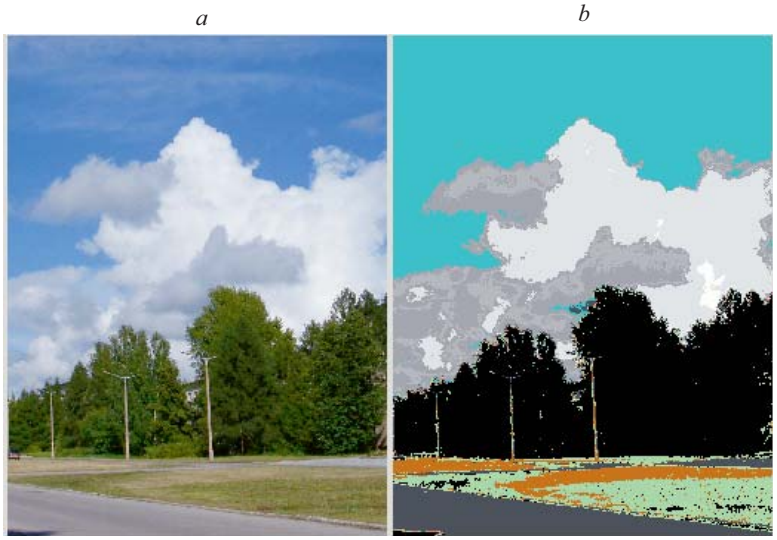

Рис. 4. Статистическое моделирование: исходное изображение (*a*), результат кластеризации (*b*)

Эксперимент 6. Использовался фрагмент снимка поврежденных сибирским шелкопрядом темнохвойных лесов южной тайги Нижнего Приангарья, полученный с помощью спутника LandSat-7. Исследуемый участок ограничен 57 и 59° северной широты и 93 и 98° восточной долготы. Обработке подвергался фрагмент размером 1001×1045 в 3-, 4- и 5-м каналах. Объем исходной выборки 1046045, параметры: $h = 7{,}5$, $N_{\min} = 0$, $T = 1{,}5$. Время обработки 9 с. Средний вес элемента 19,1. Алгоритм выделил семь классов. Результаты эксперимента показаны на рис. 6.

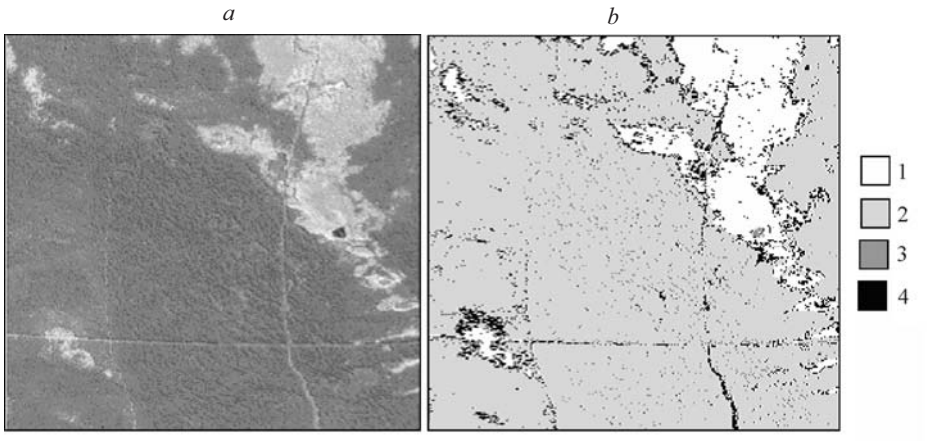

Рис. 5. Статистическое моделирование: исходный фрагмент (*a*), результат кластеризации (*b*) (1 – березовые насаждения, 2 – сосновые насаждения, 3 – вода, 4 – травянистая растительность, просеки)

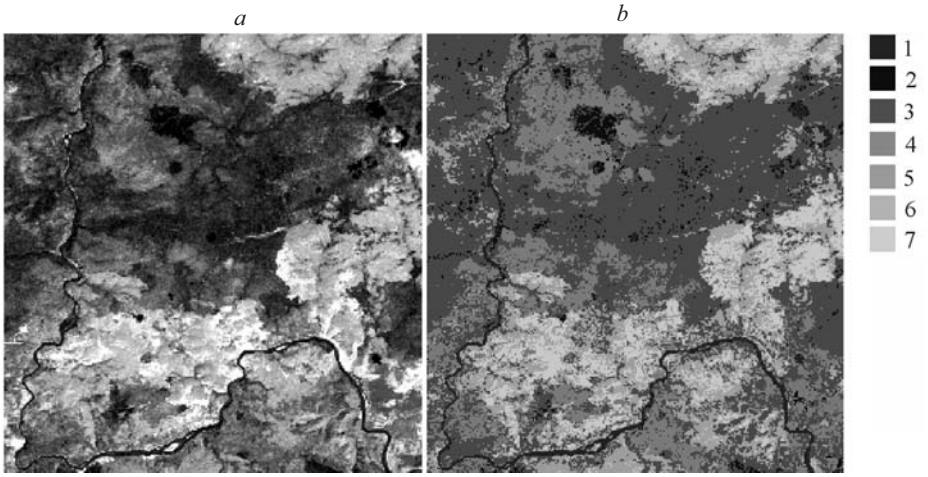

Рис. 6. Статистическое моделирование: исходный фрагмент (*a*), результат кластеризации (*b*) (1 – водная поверхность, 2 – перестойные (очень старые) березовые насаждения (возможно, вместе с темнохвойными), 3 – темнохвойные насаждения (пихтовые и еловые), 4 – погибшие от шелкопряда насаждения, 5 – спелые березовые насаждения, 6 – молодые березовые насаждения, 7 – другие нарушенные территории)

Заключение. Представленный в работе алгоритм кластеризации данных дистанционного зондирования не требует ни классифицированной обучающей выборки, ни каких-либо предположений относительно параметрической структуры данных, обеспечивая при этом высокое качество результатов. При проведении кластеризации от пользователя требуются минимальные усилия, связанные с настройкой параметров. Высокое быстродействие алгоритма обеспечивает возможность работы в диалоговом режиме. Предложенная процедура применима для обработки как космических, так и аэроснимков.

СПИСОК ЛИТЕРАТУРЫ

1. **Дейвис Ш. М., Ландгребе Д. А., Филлипс Т. Л. и др.** Дистанционное зондирование: количественный подход: Пер. с англ. /Под ред. Ф. Свейна, Ш. Дейвис. М.: Недра, 1983.
2. **Richards J. A.** Remote Sensing Digital Image Analysis: An Introduction. Berlin – Heidelberg: Springer-Verlag, 1993.
3. **Schowengerdt R. A.** Remote Sensing, Models and Methods for Image Processing. Academic Press, 1997.
4. **Halkidi M., Batistakis Y., Vazirgiannis M.** On clustering validation techniques // Journ. Intelligent Inform. Systems. 2001. **17**, N 2–3. P. 107.
5. **Berkhin P.** Survey of Clustering Data Mining Techniques. Tech. Report. Accrue Software. 2002 (http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf).
6. **Comaniciu D., Meer P.** Distribution free decomposition of multivariate data // Pattern Analys. and Appl. 1999. N 2. P. 22.
7. **Себестиан Г. С.** Процессы принятия решения при распознавании образов: Пер. с англ. Киев: Техника, 1965.

8. **Fukunaga K., Hostetler L.D.** The estimation of the gradient of a density function, with applications in patter recognition // IEEE Trans. Inform. Theory. 1975. **21**. P. 32.

9. **Загоруйко Н. Г., Елкина В. Н., Емельянов С. В., Лбов Г. С.** Пакет прикладных программ ОТЭКС. М.: Финансы и статистика, 1986.

10. **Cheng Y.** Mean shift, mode seeking, and clustering // IEEE Trans. Pattern Anal. Machine Intell. 1995. **17**. P. 790.

11. **Синявский Ю. Н., Будкина Е. А.** ППП GIPARD для автоматизированного анализа данных дистанционного зондирования // Мат. XLII Междунар. науч. студ. конф. «Студент и научно-технический прогресс». Информационные технологии. Новосибирск: Изд-во НГУ, 2004. С. 181.

Институт вычислительных технологий СО РАН,
E-mail: pestunov@ict.nsc.ru

Поступила в редакцию
7 октября 2005 г.