Построение кластерного ансамбля для сегментации гиперспектральных изображений

В. Б. Бериков[1], И. А. Пестунов[2,*]

[1]Институт математики им. С. Л. Соболева СО РАН, Новосибирск, Россия
[2]Институт вычислительных технологий СО РАН, Новосибирск, Россия
[*]Контактный e-mail: pestunov@ict.nsc.ru

Предложен алгоритм сегментации гиперспектральных изображений, основанный на коллективном подходе к кластерному анализу. Решение строится с помощью вычисления усредненной коассоциативной матрицы прототипов. Эффективность алгоритма исследуется на реальных гиперспектральных изображениях при наличии зашумленных каналов.

Ключевые слова: кластерный анализ, ансамбль алгоритмов, коассоциативная матрица, гиперспектральное изображение.

Введение

В области дистанционного зондирования Земли активно используются средства и технологии гиперспектральной съемки в видимом и ближнем инфракрасном диапазонах спектра [1]. Особенностями гиперспектральных изображений являются большое число спектральных каналов (оно может достигать нескольких сотен) и малая спектральная ширина каждого канала (порядка нескольких нанометров). Гиперспектральные изображения — это фактически трехмерные массивы данных, в них два измерения соответствуют пространственным координатам, а третье — спектральная координата, поэтому гиперспектральное изображение называют пространственно-спектральным кубом [2].

Одной из важных задач, возникающих при анализе изображений, является их сегментация, т. е. разбиение изображения на участки, однородные по какому-либо критерию [3, 4]. Сегментация гиперспектральных изображений — сложная, до конца не решенная задача. Для того чтобы выделить однородные сегменты, необходимо определить, что понимается под однородностью области изображения, сформулировать критерий, по которому можно было бы сравнивать различные варианты сегментации, а также предложить эффективный алгоритм нахождения наилучшего варианта. Имеются различные подходы к сегментации изображений. Наиболее распространенный подход к решению этой задачи основан на использовании алгоритмов кластерного анализа, применяемых к таблицам данных, сформированных из исходного изображения с привлечением разного вида признаков (спектральных, текстурных и др.) [5, 6].

Существует большое число методов кластерного анализа, различающихся способами понимания однородности, алгоритмами перебора вариантов разбиений и различными ограничениями, позволяющими учитывать специфику конкретной области [7]. Такое

разнообразие можно объяснить как существованием различных подходов к пониманию однородности кластеров и необходимостью учета специфической для каждой области дополнительной информации, так и наличием различных алгоритмов решения поставленной оптимизационной задачи.

В кластерном анализе активно развивается коллективный (ансамблевый) подход [8 – 13]. Применение этого подхода позволяет снижать зависимость результатов группировки от выбора параметров алгоритма, получать более устойчивые решения в условиях зашумленных данных. Идея построения коллективных решений, основанных на композиции простых алгоритмов, активно используется в современной теории и практике интеллектуального анализа данных, распознавания образов и прогнозирования. Коллективная решающая функция сочетает преимущества каждого из методов, используемых при построении функций. Кроме того, для каждой из базовых решающих функций может быть определена область ее наилучшей "компетентности".

Наличие различных методов кластерного анализа и вариантов оптимизационных процедур обусловливает возможность их совместного использования для формирования согласованного (или коллективного, комитетного, ансамблевого) решения. При выработке такого решения проводится группировка с "разных точек зрения" (предполагается, что эти "точки зрения" не только не противоречивы, но и дополняют друг друга; один вариант решения компенсирует "слабые стороны" других вариантов). При этом устойчивые закономерности, в соответствии с которыми формируются кластеры, взаимно "усиливаются", а неустойчивые, наоборот, "ослабляются". Ансамблевый подход делает возможным проведение распределенных вычислений (при различном местоположении переменных или объектов) в случае, когда получение общей базы данных невозможно либо требует больших затрат.

В настоящей статье развивается коллективный подход в кластерном анализе в контексте его применения к анализу гиперспектральных изображений. Одной из серьезных проблем при построении ансамблевого решения является значительная трудоемкость используемых переборных процедур. Существующие алгоритмы неспособны анализировать данные большого объема, характерные для гиперспектральных изображений. В данной работе предложен вычислительно эффективный ансамблевый алгоритм сегментации, который способен работать при большом объеме анализируемых данных. Приведены примеры обработки реальных гиперспектральных изображений при наличии зашумленных каналов.

1. Основные понятия и обозначения

В кластерном анализе требуется получить разбиение $P = \{C_1, \ldots, C_K\}$ множества некоторых элементов (объектов) $A = \{a_1, \ldots, a_N\}$ на определенное число K групп (кластеров) в соответствии с заданным критерием качества. Под критерием качества понимается некоторый функционал, зависящий от разброса внутри группы и расстояний между группами. Как правило, каждый объект описывается с помощью набора вещественных переменных $X_1, \ldots, X_n$. Через $\mathbf{x} = \mathbf{x}(a) = (x_1, \ldots, x_n)$ обозначим вектор переменных для объекта a, где $x_j = X_j(a)$, $j = 1, \ldots, n$, а через $T_{N\times n}$ — матрицу (таблицу данных) $(x(a_1), \ldots, x(a_N))^{\mathrm{T}}$. Число кластеров может быть задано или не задано; в данной работе будем считать, что требуемое число групп есть некоторый фиксированный параметр.

В анализе изображений под элементом понимается пиксель, а переменные описывают различные свойства пикселей (спектральную яркость в заданном диапазоне, тек-

стурные характеристики и т. д.). Например, RGB-изображение может быть представлено в виде таблицы данных с помощью трех переменных X_1, X_2, X_3, характеризующих интенсивность соответственно красной, зеленой и синей составляющих цвета каждого пикселя. Для гиперспектрального изображения каждый пиксель может быть охарактеризован упорядоченной последовательностью $X_1, X_2, \ldots, X_d$, где d — число спектральных каналов.

Поскольку задача поиска варианта разбиения, оптимального по заданному критерию, имеет, как правило, экспоненциальную трудоемкость, на практике чаще всего применяются приближенные итеративные алгоритмы, которые на каждом шаге проводят модификацию текущего разбиения, дающую локальное улучшение качества. Работа алгоритма управляется некоторыми параметрами или настройками алгоритма (используемыми метриками, типами межгрупповых расстояний и т. п.), задаваемыми пользователем.

2. Коллективный подход

При использовании коллективного подхода к кластерному анализу первоначально строится базовый набор вариантов группировки, по которым затем определяется итоговое разбиение на кластеры. Исходные решения формируются с использованием различных алгоритмов, по различным настройкам одного алгоритма, по случайно отобранным подсистемам переменных и т. п.

Существует несколько основных способов построения итоговых коллективных решений кластерного анализа. В первом способе от ансамбля требуют консенсуса, т. е. некоторой наилучшей степени согласованности с результатами отдельных алгоритмов. Пусть имеется L вариантов $P_1, \ldots, P_L$ разбиения множества A на кластеры. Консенсусным разбиением называют такое разбиение P^*, для которого выполняется условие

$$P^* = \arg\max_{P \in \mathcal{P}} \sum_{l=1}^{L} \varphi(P, P_l),$$

где $\mathcal{P}$ — множество всевозможных разбиений A; φ — некоторая мера сходства между двумя разбиениями. В качестве меры сходства можно использовать индекс Ранда [14].

Пусть $P_1 = \{C_{1,1}, \ldots, C_{K_1,1}\}$ и $P_2 = \{C_{1,2}, \ldots, C_{K_2,2}\}$ — два варианта группировки; $C_{k,1} = \{a_{i_1}, \ldots, a_{i_{N_{k,1}}}\}$, $C_{l,2} = \{a_{j_1}, \ldots, a_{j_{N_{l,2}}}\}$, где $N_{k,1}$ — число объектов в k-м кластере первого варианта группировки, а $N_{l,2}$ — число объектов в l-м кластере второго варианта группировки.

Индекс Ранда определяется как величина

$$\phi_R(P_1, P_2) = \frac{A + D}{G},$$

где A — число пар объектов, которые входят в одни и те же группы — в P_1 и P_2; D — число пар, которые входят в разные группы; $G = \binom{N}{2}$ — число всевозможных пар. Таким образом, данный индекс равен относительному числу правильно классифицированных (по принадлежности к кластерам) пар объектов; он принадлежит интервалу от 0 до 1; значение 1 соответствует полному согласию между двумя разбиениями.

Для поиска консенсусного разбиения часто применяются приближенные итеративные алгоритмы. Алгоритмы, реализующие данный подход, характеризуются достаточно высокой трудоемкостью.

Второе направление в теории коллективного кластерного анализа основано на вычислении коассоциативной матрицы (матрицы смежности, co-association matrix), определяющей, как часто пары объектов оказываются в одном и том же кластере в разных вариантах разбиения. Усредненная коассоциативная матрица определяется как

$$H = \frac{1}{L}\sum_{l=1}^{L} H_l,$$

где $H_l = (h_l(i,j))$ — коассоциативная матрица для l-го варианта разбиения. Элемент $h_l(i,j)$ этой матрицы равен нулю, если пара a_i и a_j $(i \neq j)$ объединена в одну группу; $h_l(i,j) = 1$, если данная пара разделена в l-м варианте разбиения по разным группам, $i,j = 1,\ldots,N$. Элементы усредненной матрицы могут рассматриваться как аналоги попарных расстояний между объектами: чем больше значение элемента, тем чаще соответствующая пара была разнесена алгоритмами, входящими в ансамбль, в разные кластеры, т. е. тем более "непохожими" являются данные объекты. Для получения итогового согласованного разбиения может быть использован алгоритм кластерного анализа, обрабатывающий таблицы попарных расстояний, на вход которого подается полученная матрица. В данной работе применяется алгоритм построения дендрограммы, в котором расстояния между группами определяются по принципу "средней связи" [3].

3. Алгоритм, основанный на матрице прототипов

Алгоритмы построения кластерного ансамбля, основанные на описываемом подходе, требуют порядка N^2 ячеек памяти для хранения элементов матрицы и характеризуются такого же порядка трудоемкостью, что делает их малоэффективными при анализе таблиц данных большой размерности. Основная идея предлагаемого в настоящей работе алгоритма основана на сочетании сжатия данных и ансамблевой группировки; в этом случае при построении коллективного решения рассматриваются не все возможные пары наблюдений, а лишь сравнительно небольшое число пар "прототипов", представляющих кластеры.

Для получения базовых вариантов группировки используется алгоритм k-средних, трудоемкость которого линейно зависит от размерности таблицы данных. Требуемое число кластеров является параметром работы алгоритма.

Под *центроидом* кластера будем понимать вектор арифметических средних всех его элементов. *Прототипом* кластера C_k назовем его элемент p_k, ближайший к центроиду данного кластера. В алгоритме CMP (Co-association Matrix of Prototypes) используется метод случайных подпространств для построения базовых элементов ансамбля. Этот алгоритм может быть использован, если исходное число переменных достаточно велико по сравнению с размерностью подпространств.

Алгоритм построения кластерного ансамбля CMP

Вход

$A = \{a_1,\ldots,a_N\}$ — множество элементов, описанное таблицей данных $T_{N\times n}$;

K — заданное число кластеров;

d_{ens} — размерность подпространства;

L — число запусков базового алгоритма;

N_{pr} — число прототипов ($N_{pr} \ll N$);

Выход

Разбиение A на K кластеров.

Начало алгоритма СМР

- Для всех $l \in \{1, \ldots, L\}$ выполняются следующие действия:
  - случайным образом выбирается d_{ens} переменных (вероятность отбора каждой переменной постоянна);
  - с помощью алгоритма k-средних множество A разбивается на N_{pr} кластеров, которым присваивают метки;
  - вычисляются прототипы кластеров и запоминаются их номера из исходной таблицы;
  - для каждого объекта из A находится ближайший к нему прототип (в текущем подпространстве) и запоминается его номер.
- Вычисляется усредненная по всем вариантам коассоциативная матрица H для всех прототипов, найденных на шаге 3.
- С помощью иерархического агломеративного алгоритма по матрице H строится разбиение C_{pr} множества прототипов на K кластеров.
- Для каждого объекта из A определяются номера ближайших прототипов по всем L базовым вариантам; с помощью голосования по большинству каждый объект относится к соответствующему кластеру из C_{pr}.

Конец алгоритма СМР.

Размерность матриц прототипов не превышает $(LN_{pr}) \times (LN_{pr})$, что намного меньше, чем $N \times N$. Трудоемкость алгоритма порядка $O\left(LN_{pr}d_{ens}N\right) + O\left((LN_{pr})^3\right)$.

4. Использование алгоритма СМР для анализа гиперспектральных изображений в условиях шумов

Для исследования эффективности работы предложенного алгоритма рассмотрено несколько тестовых гиперспектральных изображений [15]. Первое изображение (Salinas-A) размером 83×86 пикселей содержит 204 спектральных канала. Его пространственное разрешение составляет 3.7 м. На рис. 1, *а* показан RGB-композит изображения Salinas-A, полученный с использованием каналов 10, 60 и 100. На изображении представлены участки поверхности Земли с шестью типами сельскохозяйственных культур, а также участки, лишенные растительности. На рис. 1, *б* показано эталонное разбиение изображения на классы.

Для исследования поведения алгоритма в условиях шумовых искажений измерения несколько случайно выбранных каналов были заменены на значения, сгенерированные датчиком случайных чисел с равномерным распределением на интервале возможных значений. На рис. 2, *а* показаны результаты работы стандартного алгоритма k-средних на зашумленном изображении (использовались все 204 канала; зашумлены каналы 13, 25 и 148). Для анализа данного изображения использовался ансамблевый алгоритм СМР со следующими параметрами: $K = 7$, $N_{pr} = 7$, $d_{ens} = 10$, $L = 5$.

На рис. 2, *б* показано полученное с помощью алгоритма СМР разбиение (для удобства восприятия различные классы представлены разными цветами, которые не имеют отношения к цветовой схеме на рис. 1, *б*). Ансамблевый алгоритм позволил заметно улучшить качество сегментации при наличии зашумленных каналов. Время работы ансамблевого алгоритма составило 0.2 с на двухъядерном процессоре Intel Core i5 с тактовой частотой 2.8 ГГц и объемом оперативной памяти 4 Гбайт.

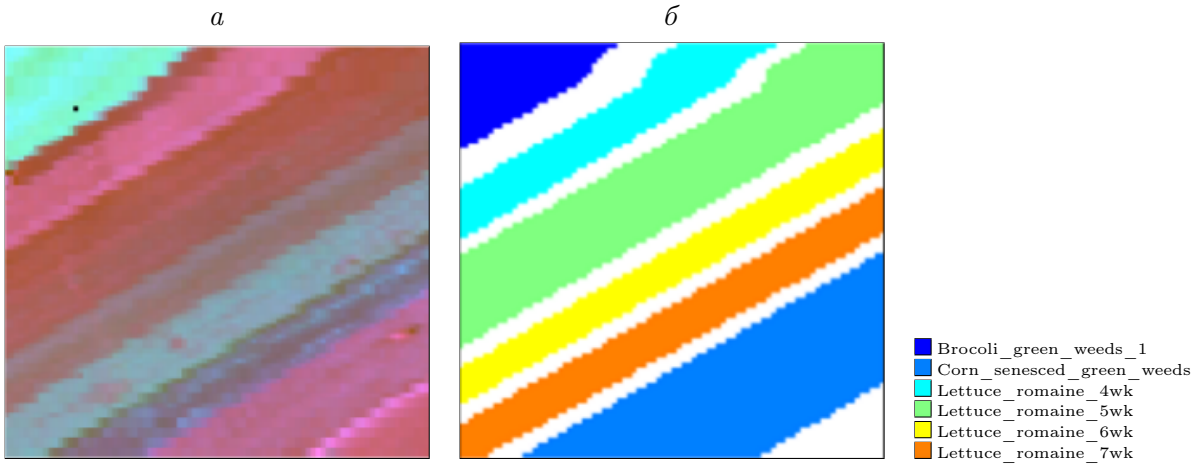

Рис. 1. Гиперспектральное изображение Salinas-A: *а* — RGB-композит (каналы 10, 60, 100); *б* — эталонное разбиение

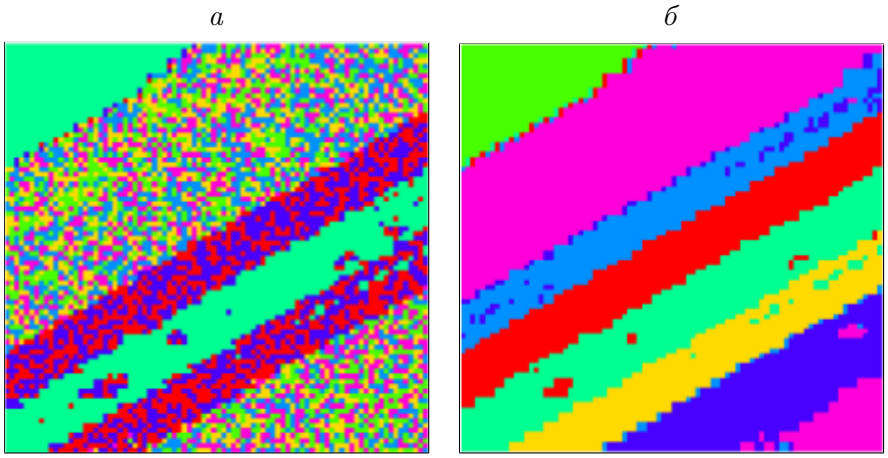

Рис. 2. Результаты работы алгоритмов на зашумленном изображении Salinas-A: *а* — алгоритм k-средних; *б* — ансамблевый алгоритм CMP

Второе изображение (Pavia University) размером 610 × 340 пикселей содержит 103 спектральных канала. Пространственное разрешение составляет 1.3 м. На рис. 3, *а* показан RGB-композит изображения (каналы 40, 50 и 70), а на рис. 3, *б* приведено эталонное разбиение изображения на тематические классы.

На рис. 4, *а* показаны результаты работы стандартного алгоритма k-средних на зашумленном изображении (использовались все 103 канала; искажены каналы 13, 25 и 148). На рис. 4, *б* показаны результаты сегментации этого же изображения алгоритмом CMP с параметрами: $K = 10$, $N_{pr} = 10$, $d_{ens} = 7$, $L = 5$. Хотя полученная сегментация и отличается от эталонной, ансамблевый алгоритм дал намного более четкие представления объектов снимка. Время работы ансамблевого алгоритма CMP на данном изображении составило 27.7 с. Заметим, что в данном случае применение ансамблевого алгоритма без использования прототипов (рассматриваются все возможные пары наблюдений) затруднительно, поскольку для хранения элементов коассоциативной матрицы потребуется порядка 10^{10} ячеек оперативной памяти; время обработки также возрастет на несколько порядков.

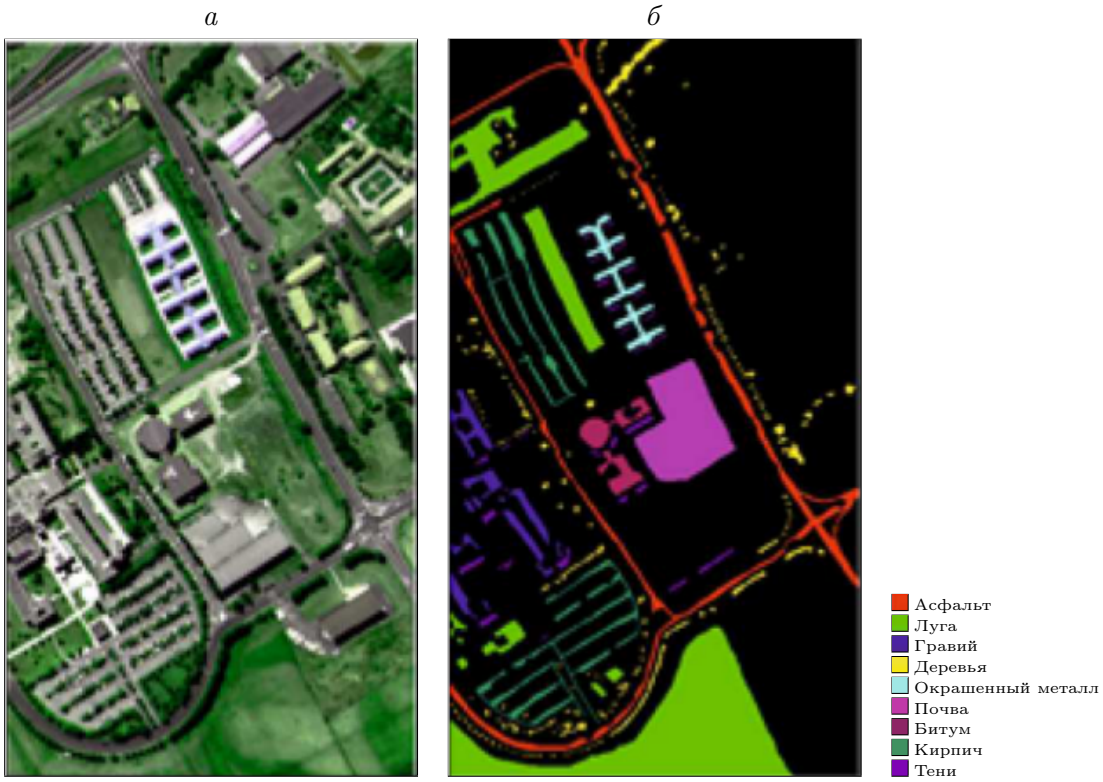

Рис. 3. Гиперспектральное изображение Pavia University: *а* — RGB-композит (каналы 40, 50, 70); *б* — эталонное разбиение на классы

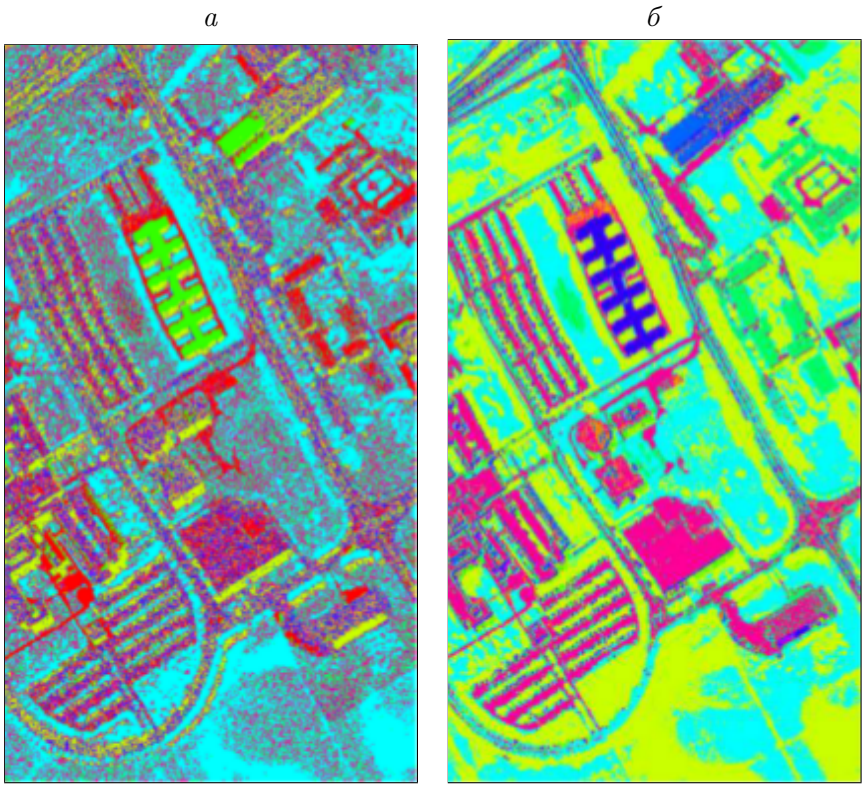

Рис. 4. Результаты работы алгоритмов на зашумленном изображении Pavia University: *а* — алгоритм k-средних; *б* — ансамблевый алгоритм CMP

Таким образом, предложен ансамблевый алгоритм сегментации гиперспектральных изображений СМР, основанный на коассоциативных матрицах прототипов. В отличие от существующих алгоритмов сегментации предложенный алгоритм позволяет повышать устойчивость результатов анализа в условиях шумовых искажений и увеличивать быстродействие при обработке данных достаточно большого объема. В дальнейшем планируется развитие алгоритма: привлечение не только спектральных, но и контекстных признаков, учет индексов качества и меры разнообразия вариантов группировки при формировании итогового решения.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (гранты № 14-07-00249-а, № 13-07-12202-офи_м).

Список литературы / References

[1] **Бондур В.Г.** Современные подходы к обработке больших потоков гиперспектральной и многоспектральной аэрокосмической информации // Исследование Земли из космоса. 2014. № 1. С. 4–16.
Bondur, V.G. Modern approaches for processing of big hyperspectral aerospace data // Earth Observation and Remote Sensing. 2014. No. 1. P. 4–16. (In Russ.)

[2] **Шовенгердт Р.А.** Дистанционное зондирование. Модели и методы обработки изображений. М.: Техносфера, 2010. 560 с.
Schowengerdt, R.A. Remote sensing: models and methods for image processing. New York: Acad. Press, 2006. 560 p.

[3] **Duda, R.O., Hart, P.E., Stork, D.G.** Pattern classification. Second edition. New York: Wiley, 2000. 680 p.

[4] **Гонсалес Р., Вудс М.** Цифровая обработка изображений. М.: Техносфера, 2012. 1104 с.
Gonzales, R., Woods, R. Digital image processing. Third edition. New Jersey: Pearson Education Inc., Prentice Hall, 2008. 954 p.

[5] **Пестунов И.А., Синявский Ю.Н.** Алгоритмы кластеризации в задачах сегментации спутниковых изображений // Вест. КемГУ. 2012. Т. 52, № 4/2. С. 110–125.
Pestunov, I.A., Sinyavskiy, Yu.N. Clustering algorithms in satellite images segmentation tasks // Bulletin of KemSU. 2012. Vol. 52, No. 4/2. P. 110–125. (In Russ.)

[6] **Пестунов И.А., Рылов С.А.** Алгоритмы спектрально-текстурной сегментации спутниковых изображений высокого пространственного разрешения // Вест. КемГУ. 2012. Т. 52, № 4/2. С. 104–110.
Pestunov, I.A., Rylov, S.A. Spectral-textural segmentation algorithms for satellite images with high spatial resolution // Bulletin of KemSU. 2012. Vol. 52, No. 4/2. P. 104–110. (In Russ.)

[7] **Jain, A.K.** Data clustering: 50 years beyond K-means // Patt. Recogn. Lett. 2010. Vol. 31(8). P. 651–666.

[8] **Topchy, A., Law, M., Jain, A., Fred, A.** Analysis of consensus partition in cluster ensemble // Fourth IEEE Intern. Conf. on Data Mining (ICDM'04). 2004. P. 225–232.

[9] **Ghosh, J., Acharya, A.** Cluster ensembles // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. Vol. 1(5). P. 305–315.

[10] **Пестунов И.А., Бериков В.Б., Синявский Ю.Н.** Сегментация многоспектральных изображений на основе ансамбля непараметрических алгоритмов кластеризации // Вест. СибГАУ. 2010. Вып. 5 (31). С. 56–64.

Pestunov, I.A., Berikov, V.B., Sinyavskiy, Yu.N. Algorithm for multispectral image segmentation based on ensemble of nonparametric clustering algorithms // Vestnik SibGAU. 2010. Vol. 5 (31). P. 56–64. (In Russ.)

[11] **Пестунов И.А., Бериков В.Б., Куликова Е.А., Рылов С.А.** Ансамблевый алгоритм кластеризации больших массивов данных // Автометрия. 2011. Т. 47, № 3. С. 49–58.
Pestunov, I.A., Berikov, V.B., Kulikova, E.A., Rylov, S.A. Ensemble of clustering algorithm for large datasets // Optoelectronics, Instrumentation and Data Processing. 2011. Vol. 47, iss. 3. P. 245–252.

[12] **Пестунов И.А., Рылов С.А., Бериков В.Б.** Иерархические алгоритмы кластеризации для сегментации мультиспектральных изображений // Автометрия. 2015. Т. 51, № 4. С. 12–22.
Pestunov, I.A., Rylov, S.A., Berikov, V.B. Hierarchical clustering algorithms for segmentation of multispectral images // Optoelectronics, Instrumentation and Data Processing. 2015. Vol. 51, iss. 4. P. 329–338.

[13] **Berikov, V.** Weighted ensemble of algorithms for complex data clustering // Patt. Recogn. Lett. 2014. Vol. 38. P. 99–106.

[14] **Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J., Perona, I.** An extensive comparative study of cluster validity indices // Patt. Recogn. Lett. 2013. Vol. 46, iss. 1. P. 243–256.

[15] Hyperspectral Remote Sensing Scenes. Available at: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (accessed 20.11.2015).

Поступила в редакцию 8 декабря 2015 г.,
с доработки — 12 января 2016 г.

Creating a cluster ensemble for hyperspectral images segmentation

Berikov, Vladimir B.[1], Pestunov, Igor A.[2,*]

[1]Sobolev Institute of Mathematics SB RAS, Novosibirsk, 630090, Russia
[2]Institute of Computational Technologies SB RAS, Novosibirsk, 630090, Russia
[*]Corresponding author: Pestunov, Igor A., e-mail: pestunov@ict.sbras.ru

Ensemble approach has been actively developed in cluster analysis. This approach helps to reduce the dependence of the results on the choice of the algorithm parameters and to receive more stable solutions for noisy data.

In this work we suggest an algorithm of hyperspectral images segmentation based on the ensemble clustering. For this purpose we consider a method of solution formation using co-association matrices that define how often pairs of objects appear in the same cluster in different variants of partitioning.

One of the serious problems in constructing the ensemble solution is considerable running time of algorithms and necessity to store co-association matrices of large dimension in memory. Existing algorithms are not able to analyze large amounts of data, typical of hyperspectral images. In this paper we describe a computationally efficient algorithm for clustering ensemble. The main idea of the algorithm is based on the combination of data compression and ensemble clustering. While constructing

ensemble solution one should examine not all pairs of observations, but rather only small number of pairs of "prototypes" that represent clusters.

The effectiveness of the algorithm is illustrated on real hyperspectral images in the presence of noisy channels. It is shown that the proposed algorithm can improve the quality for the results of the noisy data analysis to handle a large images.

Keywords: cluster analysis, ensemble algorithms, co-association matrix, hyperspectral image.

Acknowledgements. This research was partially supported by RFBR (grant No. 14-07-00249, No. 13-07-12202).

Received 8 December 2015
Received in revised form 12 January 2016