УДК 528.852

КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ*

В. В. Асмус[1], А. А. Бучнев[2], В. П. Пяткин[2]

[1] *Государственное учреждение «Научно-исследовательский центр "Планета"», 123242, Москва, Б. Предтеченский пер., 7*
[2] *Учреждение Российской академии наук Институт вычислительной математики и математической геофизики Сибирского отделения РАН, 630090, г. Новосибирск, просп. Академика Лаврентьева, 6*
E-mail: pvp@ooi.sscc.ru

Рассматривается система кластерного анализа (неконтролируемой классификации) данных дистанционного зондирования Земли. Система представлена тремя методами: K-средних, анализа мод многомерных гистограмм и гибридным, объединяющим метод анализа мод многомерных гистограмм с их последующей иерархической группировкой.

Ключевые слова: дистанционное зондирование, распознавание образов, кластерный анализ.

Введение. В предлагаемой работе рассматриваются функциональные возможности системы кластерного анализа (неконтролируемой классификации) в программном комплексе обработки данных дистанционного зондирования Земли (ДДЗЗ), созданном в Научно-исследовательском центре «Планета» и Институте вычислительной математики и математической геофизики СО РАН. Это логическое продолжение работы [1], где описана система контролируемой классификации ДДЗЗ. Характеризуя методы кластеризации в целом, следует отметить, что они отыскивают в данных не те структуры, которые там реально существуют, а те, для поиска которых предназначены [2]. Поэтому надёжность результатов кластеризации часто можно оценить лишь сравнением нескольких вариантов обработки. Система кластерного анализа в программном комплексе представлена тремя методами: K-средних, анализа мод многомерной гистограммы [2] и гибридным, объединяющим метод анализа мод многомерной гистограммы с их последующей иерархической группировкой.

Метод K-средних основан на итеративной процедуре отнесения векторов признаков к классам по критерию минимума расстояния от вектора до центра класса. Оптимальным считается такое разбиение входных векторов на кластеры, при котором внутриклассовый разброс не может быть уменьшен при переносе какого-либо вектора из одного кластера в другой.

Алгоритм K-средних согласно [3] является одним из способов, называемых «методами центра тяжести» и используемых в задачах автоматической классификации данных, и представляет собой вариант метода динамических сгущений.

Как правило, значение K (число кластеров в наборе данных) неизвестно и должно извлекаться исключительно из самих данных [4]. В то же время во многих кластерных алгоритмах значение K является входным параметром, и очевидно, что качество получаемых кластеров в большой степени зависит от оценки K. Разбиение данных на слишком

*Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (проект № 10-07-00131).

большое количество кластеров усложняет результат и делает трудными его интерпретацию и анализ, однако разбиение на слишком малое количество кластеров приводит к потере информации. Некоторые авторы называют проблему определения числа кластеров «фундаментальной проблемой правильности кластеризации». Предприняты многочисленные попытки оценить правильное значение K (см., например, [5]), но сложность соответствующих алгоритмов ограничивает их применение небольшими наборами данных (ДДЗЗ к таковым не относятся). В реализациях рассматриваемого программного комплекса алгоритмов кластеризации значение K является входным параметром. Оно задаётся экспертом на основе эвристического анализа исходного набора векторов измерений, представленного в виде изображения.

Выбор алгоритма. Для алгоритма K-средних существует два способа пересчёта центров кластеров при выполнении итерационных операций. В соответствии с первым методом (метод Ллойда [6]) выполнение каждой итерации алгоритма состоит в распределении всех векторов данных по кластерам исходя из минимума расстояния до центров кластеров и последующем пересчёте центров кластеров согласно полученному распределению. В варианте, предложенном Мак-Квином [7], всякий раз, когда выясняется, что вектор, находящийся в j-м кластере, на самом деле ближе к центру k-го кластера, этот вектор переводится из кластера j в кластер k с пересчётом центров и объёмов этих кластеров.

В обоих вариантах количество выполняемых алгоритмами итераций ограничивается значением параметра Iterations. Дополнительным параметром управления временем работы алгоритма является значение Delta — точность вычислений. В первом варианте алгоритм заканчивает работу на итерации с номером k, если $|E_{k-1} - E_k| \leq \mathrm{Delta}$, где E_k — сумма квадратов расстояний (ошибок) всех векторов до центров соответствующих кластеров на k-й итерации. Во втором варианте алгоритм заканчивает работу на очередной итерации, если число векторов, переведённых из одного кластера в другой, не превосходит заданной величины V_Transp.

Входные данные. В программном комплексе процедуре кластеризации могут подвергаться как исходный набор векторов признаков, так и некоторый набор векторов, каждый из которых является «представителем» группы векторов из исходного набора данных. В последнем случае этап собственно кластеризации предваряется построением группы представителей. В результате исходному вектору приписывается номер кластера, в который попал представитель этого вектора. Цель такого подхода состоит в сокращении времени работы алгоритма за счёт уменьшения количества обрабатываемых векторов (естественно, с некоторым ухудшением качества кластеризации).

Алгоритм кластеризации по представителям состоит в выполнении следующих шагов:

1. На основе заданного соотношения α производится разделение векторов на чистые и смешанные. Под смешанными понимаем векторы, компоненты которых либо формируются за счёт попадания в поле зрения съёмочной аппаратуры нескольких объектов, либо искажены влиянием фона.

Вначале для исходного набора векторов измерений с помощью выбранного многомерного оператора градиента Grad (оператор Робертса — Превитта — Собела) рассчитывается градиентное изображение и одновременно строится гистограмма градиентов. Значение многомерного оператора Grad в позиции (x, y) многоспектрального изображения определяется следующим образом:

$$\mathrm{Grad}(x, y) = \sum_{i=1}^{n} \mathrm{Grad}^i(x, y).$$

Здесь и далее n — количество спектральных диапазонов, $\mathrm{Grad}^i(x, y)$ — значение гради-

ентного оператора в позиции (x, y) i-го спектрального диапазона:

$$\mathrm{Grad}^i(x, y) = \sqrt{G_x^2 + G_y^2},$$

где G_x и G_y — частные производные по x и y соответственно.

Известно [8], что операция вычисления градиента в приведённом виде не совсем корректна из-за того, что она применима к скалярным функциям и неприменима к векторным, для которых разработаны различные обобщения понятия градиента. Использование этих обобщений для задачи кластеризации в данной работе является неоправданным из-за значительных вычислительных трудностей.

Исходя из заданного α по гистограмме градиентов определяется порог, разделяющий векторы на смешанные и чистые.

2. Объединение чистых векторов в связные компоненты. На этом этапе все чистые векторы объединяются в связные компоненты, которые последовательно нумеруются. Соответствующий алгоритм, идейно близкий к алгоритму заполнения областей с произвольной границей по критерию связности [7], может выделять и нумеровать одновременно любое количество многосвязных областей без ограничений на их форму и ширину контуров. Для каждой связной компоненты вычисляется вектор средних, который и является представителем группы векторов, входящих в эту компоненту.

3. Итеративная кластеризация векторов средних одним из вариантов алгоритма K-средних.

4. Распределение связных компонент по кластерам. На этом этапе связные компоненты получают новые номера. Новый номер присваивается компоненте в соответствии с номером кластера, в который попал вектор средних этой компоненты.

5. Формирование кластерного образа с одновременной кластеризацией смешанных векторов, выделенных в п. 1. Кластеризация смешанных векторов производится по принципу минимума расстояния до центров кластеров $C_1, \ldots, C_k$. Смешанный вектор z будет отнесён к ближайшему кластеру ω_i, если

$$\|C_i - z\| < 0{,}5 \max \|C_i - C_l\|, \quad i, l = 1, \ldots, k, \quad i \neq l.$$

Заметим, что процесс разбиения векторов на чистые и смешанные является в большой степени эвристическим: незначительные изменения параметра, задающего соотношение чистых и смешанных векторов, могут приводить к немалым изменениям получаемых результатов.

Используемые метрики. В процессе работы алгоритмов расстояние между векторами x и y определяется на основе одной из трёх метрик (норм):

— метрика Евклида (L_2-норма)

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

— метрика сити-блок (L_1-норма)

$$\rho(x, y) = \sum_{i=1}^{n} |x_i - y_i|,$$

— метрика Чебышева (L_∞-норма)

$$\rho(x, y) = \max |x_i - y_i|, \quad i = 1, \ldots, n.$$

Здесь n — размерность векторов (количество спектральных диапазонов).

Выбор метрики определяет форму получаемых кластеров. Для метрики Евклида эквидистантными поверхностями являются гиперсферы, центры которых совпадают с центрами кластеров.

Для L_1-нормы эквидистантными являются поверхности выпуклых гипермногогранников (гипероктаэдров); в случае $n = 3$ для кластера с центром $x^c = (x_1^c, x_2^c, x_3^c)$ точки $x = (x_1, x_2, x_3)$ эквидистантной поверхности должны удовлетворять уравнению

$$\rho(x, x^c) = |x_1 - x_1^c| + |x_2 - x_2^c| + |x_3 - x_3^c| = R, \quad R = \text{const},$$

определяющему октаэдр с центром в точке x^c и координатами вершин: $(x_1^c - R,\ x_2^c,\ x_3^c)$, $(x_1^c + R,\ x_2^c,\ x_3^c)$, $(x_1^c,\ x_2^c - R,\ x_3^c)$, $(x_1^c,\ x_2^c + R,\ x_3^c)$, $(x_1^c,\ x_2^c,\ x_3^c - R)$, $(x_1^c,\ x_2^c,\ x_3^c + R)$.

Для метрики Чебышева эквидистантными являются поверхности гиперкубов, центры которых находятся в центрах кластеров.

Выбор начальных центров кластеров. Известно [9], что результаты кластеризации методом K-средних зависят от выбора начальных центров кластеров, а в некоторых случаях даже от порядка расположения входных данных. В программном комплексе реализованы следующие варианты выбора начальных центров кластеров.

1. Пусть m — вектор средних исходной выборки. Вычисляется значение квадратного корня из суммы дисперсий в каналах:

$$\sigma = \Big(\sum_i \sigma_i^2 \Big)^{1/2},$$

где σ_i^2 — дисперсия в i-м канале, а также значение переменной $Ak = \text{C\_Sdisp}\sigma$ (C_Sdisp — входной параметр программы, C_Sdisp $\geq$ 0,05). В качестве первого начального центра кластера берётся первый вектор. Затем, если расстояние от очередного вектора до ближайшего центра кластера больше Ak, этот вектор образует центр нового кластера; в противном случае, если выбран алгоритм Мак-Квина, вектор присоединяется к ближайшему кластеру. Как только получится требуемое число центров кластеров, остальные векторы относятся к ближайшим кластерам. Заметим, что при таком выборе начальных центров количество кластеров может оказаться меньше требуемого (если не найдётся нужного количества векторов, отстоящих друг от друга на расстояние, большее Ak). Кроме того, именно при данном выборе начальных центров результат кластеризации зависит от порядка векторов в исходной выборке.

2. По полученным значениям компонент среднего вектора m и стандартным отклонениям формируются векторы v_b и v_e с компонентами

$$v\_b = (m_1 - \sigma_1, m_2 - \sigma_2, \ldots, m_n - \sigma_n), \quad v\_e = (m_1 + \sigma_1, m_2 + \sigma_2, \ldots, m_n + \sigma_n)$$

(n — количество каналов). Центры кластеров распределяются равномерно вдоль вектора, соединяющего v_b и v_e, т. е. для i-го кластера центр находится в позиции с координатами $v\_i = v\_b + idv$, $i = 0, 1, \ldots, K - 1$, где dv — вектор с координатами $dv = 2/(K - 1)(\sigma_1, \sigma_2, \ldots, \sigma_n)$. Затем все векторы распределяются по кластерам в соответствии с критерием близости к центрам. По этой схеме определяются начальные центры кластеров в программном комплексе Erdas Imagine. При таком выборе начальных центров количество кластеров может быть меньше требуемого (для некоторых центров может не оказаться векторов, наиболее близких к ним).

3. Случайное распределение исходного набора векторов по K кластерам: очередной вектор исходной выборки помещается в кластер, номер которого выдаётся генератором

случайных чисел. После просмотра всей выборки определяются начальные центры кластеров. Как отмечается в [8], это наиболее часто используемый метод инициализации центров кластеров.

Отбор векторов для кластеризации. Результаты кластеризации методом K-средних в значительной степени зависят от дисперсии входных данных [10]: большая дисперсия стремится нарушить форму получаемых кластеров. В связи с этим предусмотрена возможность ограничения набора векторов для кластеризации: обрабатываются только те векторы, которые не выходят за границу эквидистантной поверхности (в случае евклидовой метрики это гиперсфера) $\rho(x, y) = \text{Distance\_M}\sigma$ с центром в векторе средних m (векторы, не удовлетворяющие этому условию, назовём «далёкими»). Distance_M — параметр программы. По окончании процесса кластеризации далёкие векторы в зависимости от значений некоторых параметров могут быть полностью либо частично распределены по кластерам на основе минимума расстояния до центра кластеров.

Кроме того, на исходном изображении могут присутствовать объекты, фактически являющиеся шумом по отношению к интересующей эксперта части изображения (например, таким объектом при анализе прибрежных водных акваторий является суша). В связи с этим для исключения из процесса обработки ненужных объектов предусмотрен механизм маскирования векторов изображения: с обрабатываемым изображением связывается одноканальное изображение, в котором пикселы со значением 255 разрешают обработку соответствующих векторов исходного изображения (физические размеры обоих изображений должны быть одинаковыми). Влияние маски на результат кластеризации демонстрируют рис. 1–3. На рис. 1 показано исходное изображение части прибрежной акватории Чёрного моря. Космический снимок получен с искусственного спутника Земли "Aqua" 23 мая 2006 г. (сканер MODIS, разрешение 250 м, спектральные каналы 0,620–0,670; 0,545–0,565; 0,459–0,479 мкм). На рис. 2 приведены результаты кластеризации всего изображения алгоритмом Ллойда (выделялось 10 кластеров), а рис. 3 демонстрирует влияние маскирования суши и облачности на результат кластеризации.

2. Метод анализа мод многомерной гистограммы. Алгоритм K-средних может быть отнесён к классу параметрических, так как он неявным образом предполагает природу плотности вероятности: кластеры стремятся иметь конкретную геометрическую форму, зависящую от выбранной метрики. Альтернативой является подход, основанный

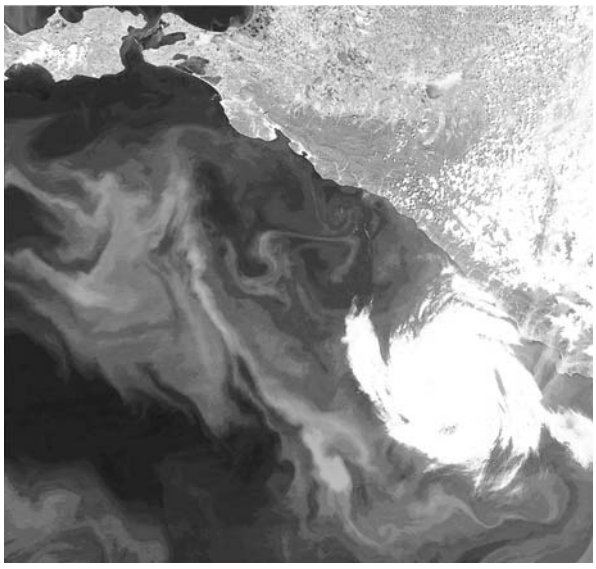

Рис. 1

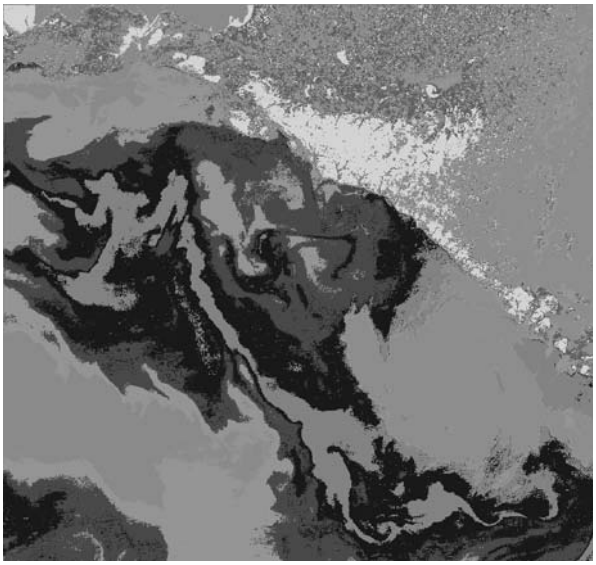

Рис. 2

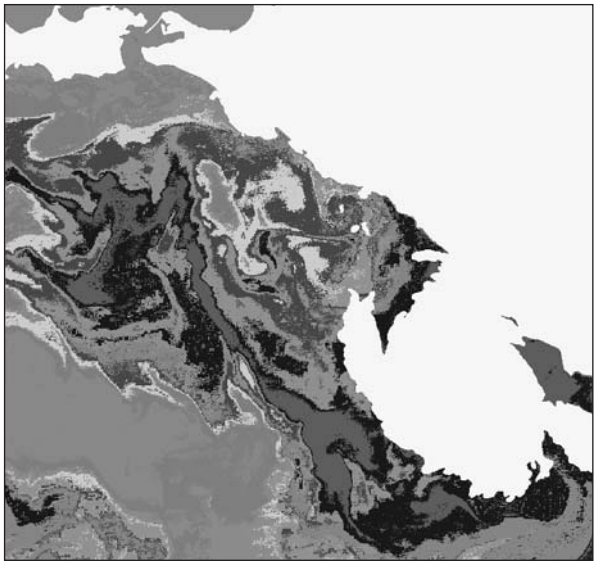

Рис. 3

на предположении, что исходные данные есть выборка из многомодового закона распределения, причём векторы, отвечающие отдельной моде, образуют кластер [11, 12]. Таким образом, задача сводится к анализу мод многомерных гистограмм.

В программный комплекс включена реализация метода, описание основных шагов которого приведено в [2, 13].

Гистограмма генерируется последовательным просмотром векторов данных и сравнением каждого вектора с текущим списком векторов. При этом либо изменяется соответствующее значение элемента гистограммы (частоты вектора), либо вектор добавляется в список. Для вычисления адресов векторов в списке используется хэш-кодирование. Первым шагом модального анализа является поиск ближайших соседей данного вектора списка среди других векторов списка. По определению вектор x есть ближайший сосед вектора y, если $|x_i - y_i| \leq 1$ для $i = 1, \ldots, n$. Каждый из возможных ближайших соседей данного вектора x может быть получен из него прибавлением вектора сдвига, компоненты которого принимают значения из множества $\{-1, 0, 1\}$. Алгоритмически i-й вектор сдвига, $i = 1, 2, \ldots, 3^n - 1$, можно получить, уменьшив на 1 каждый из коэффициентов представления числа i в троичной системе счисления. Поскольку в реальной гистограмме присутствуют далеко не все ближайшие соседи, то для эффективного их поиска векторы предварительно упорядочиваются в многомерные бинарные деревья. В этом случае время поиска всех ближайших соседей данного вектора становится пропорциональным числу реально существующих соседей. При построении дерева векторы x рассматриваются как n-мерные ключи. Вначале рассчитываются дисперсии по всем координатам векторов и определяется координата j, имеющая максимальную дисперсию. Медианное значение выборки по этой координате используется в качестве ключа для разделения множества векторов на два подмножества: в одно подмножество помещаются векторы, значение которых по координате j меньше порогового значения, а в другое — векторы, у которых значение координаты превосходит порог. Каждое из полученных подмножеств делится затем аналогичным образом.

Далее проводится локализация мод гистограммы. Каждому вектору на основе анализа его ближайших соседей ставится в соответствие градиент и приписывается номер вектора с максимальным значением градиента. Если градиент меньше 0, то это означает, что координаты вектора являются координатами локального максимума и вектору приписывается

Рис. 4

его собственный номер. В итоге каждая мода гистограммы сопоставляется с ориентированным графом, корень которого соответствует точке моды. Если количество получаемых кластеров (количество локальных максимумов гистограммы) больше заданного порога, то проводится сглаживание гистограммы. Сглаживание осуществляется либо путём замены частоты $h(x)$ вектора x средним значением частот его ближайших соседей, либо путём уменьшения «разрешения» векторов данных, т. е. делением компонент векторов на 2.

На завершающем этапе выполняется раскраска ориентированного графа одним цветом (рис. 4), т. е. всем вершинам графа присваивается значение, которое назначено его корню. Космический снимок получен с искусственного спутника Земли «Метеор-3М» 29 мая 2004 г. (сканер МСУ-Э, разрешение 40 м, спектральные каналы 0,5–0,6; 0,6–0,7; 0,8–0,9 мкм).

На рис. 5 приведён результат кластеризации исходного изображения, представленного на рис. 4, описанным методом. Выделено 15 кластеров.

Гибридный метод: анализ мод многомерной гистограммы с их последующей иерархической группировкой. Практическое использование метода анализа мод многомерной гистограммы показывает, что зачастую получение приемлемого результата является весьма трудоёмким процессом и требует высокой квалификации эксперта-исследователя. Причина этого, вероятно, в том, что алгоритм многопараметрический (в частности, на решение оказывает большое влияние способ сглаживания гистограммы).

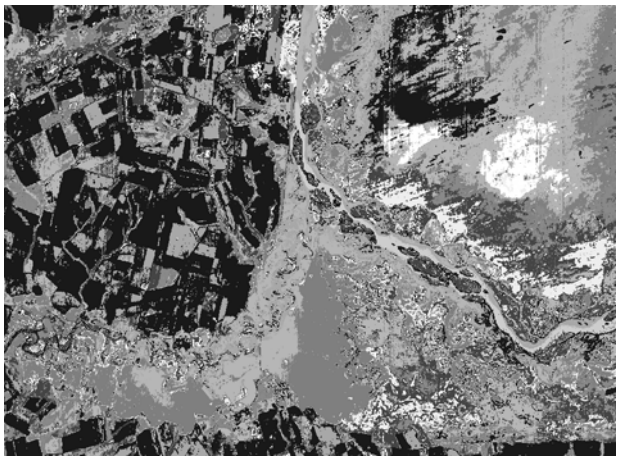

Рис. 5

В связи с этим система кластеризации дополнена двухэтапной процедурой (с сохранением всех ранее существовавших функций): на первом этапе выполняется предварительное разбиение исходной выборки на кластеры с помощью модального анализа, а на втором — для получения окончательного результата используется иерархическая группировка [14]. Заметим, что её применение для кластеризации исходного набора векторов нереально из-за того, что в алгоритме матрица расстояний состоит (в начале работы алгоритма) из $N(N-1)/2$ элемента, где N — количество векторов. Предварительное использование модального анализа позволяет сократить объём данных до разумных пределов. В качестве входных данных для иерархической группировки можно взять векторы средних группы векторов, связанных с каждой модой многомерной гистограммы. Напомним, что на каждом шаге восходящей иерархической классификации объединяются два кластера, расстояние между которыми минимально. Среди всех возможных расстояний между кластерами [10, 14] для ускорения вычислений берётся простейшее — расстояние между векторами средних в кластерах.

Достоинством иерархической группировки является то, что после построения иерархического дерева кластеризации его можно «разреза́ть» на любом уровне иерархии, т. е. получать разные кластерные карты, не запуская снова процесс кластеризации.

На рис. 6 приведён результат кластеризации исходного изображения, представленного на рис. 4, описанным методом. Выделено 15 кластеров.

Выходные данные. Последним этапом работы всех алгоритмов является сортировка полученных кластеров по убыванию их объёмов и подсчёт соответствующих статистик: объёмов, векторов средних и девиаций (стандартных отклонений) в каналах для каждого кластера. Эти результаты записываются при необходимости в файл на диске. Туда же включается число векторов данных, не вошедших ни в один из кластеров, т. е. попавших в $(K+1)$-й кластер. Эти данные будут основой для анализа разделимости полученных кластеров.

Результатом функционирования классификаторов в рабочем режиме будет одноканальное (байтовое) изображение, значениями пикселов которого являются номера кластеров. Это изображение окрашивается в предопределённые цвета, которые в интерактивном режиме могут быть заменены цветами, определяемыми пользователем. Изображение сохраняется на диске в виде стандартного BMP-файла.

К выходному изображению можно применить функцию постклассификации для удаления изолированных пикселов (генерализация данных). Эта функция работает в двух режимах: Vote, при котором центральный пиксел окрестности 3×3 заменяется макси-

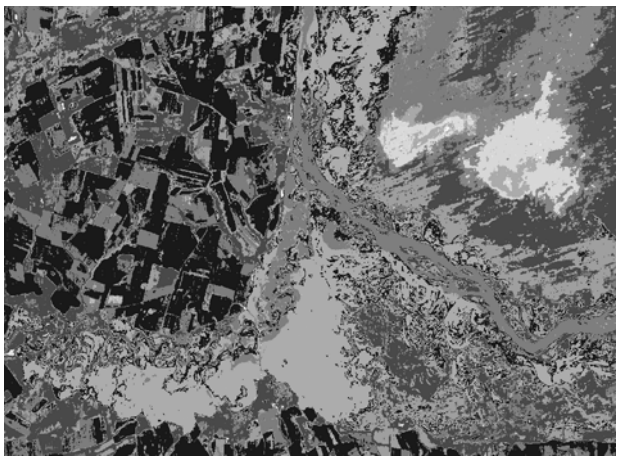

Рис. 6

мумом гистограммы окрестности, и Allsame, при котором центральный пиксел такой же окрестности меняется, когда все окружающие его пикселы имеют одинаковое значение.

Заключение. Представленная в данной работе система кластерного анализа ДДЗЗ в течение ряда лет используется в Научно-исследовательском центре «Планета» Роскомгидромета при решении широкого круга научных и практических задач. Возможность сравнения различных вариантов получаемой классификации позволяет выбрать решение, наиболее адекватно отражающее внутреннюю структуру данных.

СПИСОК ЛИТЕРАТУРЫ

1. **Асмус В. В., Бучнев А. А., Пяткин В. П.** Контролируемая классификация данных дистанционного зондирования Земли // Автометрия. 2008. **44**, № 4. С. 60–67.
2. **Асмус В. В., Вадас В., Карасев А. Б. и др.** Программный комплекс классификации многозональных данных // Исседование Земли из космоса. 1988. № 3. С. 86–94.
3. **Дидэ Э.** Методы анализа данных: Пер. с фр. М.: Финансы и статистика, 1985. 357 с.
4. **Xu R., Wunsh D. II.** Survey of clustering algorithms // IEEE Trans. Neural Networks. 2005. **16**, N 3. P. 645–678.
5. **Kothari R., Pitts D.** On finding the number of clusters // Patt. Recogn. Lett. 1999. **20**, N 4. P. 405–416.
6. **Lloyd S. P.** Least squares quantization in PCM // IEEE Trans. Inform. Theory. 1982. **28**, N 2. P. 129–137.
7. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. of the 5th Berkeley Symposium on Mathematical Statistical and Probability. 1967. Vol. 1. P. 281–297.
8. **Гонсалес Р., Вудс Р.** Цифровая обработка изображений. М.: Техносфера, 2005. 1072 с.
9. **Pena J. M., Lozano J. A., Larranaga P.** An empirical comparison of four initialization methods for the k-means algorithm // Patt. Recogn. Lett. 1999. **20**, N 10. P. 1027–1040.
10. **Marques de Sa J. P.** Pattern recognition: concepts, methods, and applications. N. Y.: Springer-Verlag, 2001. 328 p.
11. **Narendra P. M., Goldberg M.** A non-parametric clustering scheme for landsat // Patt. Recogn. 1977. **9**, N 4. P. 207–215.
12. **Красиков В. А., Шамис В. А.** Кластерная процедура на базе многомерной гистограммы распределения // Исследование Земли из космоса. 1982. № 2. С. 107–114.
13. **Асмус В. В., Бучнев А. А., Пяткин В. П.** Программный комплекс для обработки данных дистанционного зондирования Земли // Тр. XXXII Междунар. конф. «Информационные технологии в науке, образовании, телекоммуникации и бизнесе» (IT+SE'2005). Запорожье: ЗГУ, 2005. С. 229–232.
14. **Жамбю М.** Иерархический кластер-анализ и соответствия: Пер. с фр. М.: Финансы и статистика, 1988. 342 с.

Поступила в редакцию 29 октября 2009 г.