РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2007, том 43, № 1

УДК 528.852

ОЦЕНКА КАЧЕСТВА КЛАССИФИКАЦИИ МНОГОСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ ГИСТОГРАММНЫМ МЕТОДОМ *

В. С. Сидорова

Институт вычислительной математики и математической геофизики СО РАН, г. Новосибирск
E-mail: svs@ooi.sscc.ru

Предложена мера оценки качества неконтролируемой классификации спутниковых изображений кластерным методом, основанным на многомерной гистограмме. Мера используется при выборе лучших распределений векторов, соответствующих различной детальности представления данных. Экспериментально показано, что хорошо разделенные унимодальные кластеры распределений соответствуют представительным информационным классам зондируемой поверхности Земли.

Введение. Одной из задач кластерного анализа является проверка качества полученных распределений векторов. Под качеством обычно понимается хорошее разделение кластеров. Кластерные алгоритмы группируют векторы по заданным критериям, правилам, но после классификации остаются вопросы: какое распределение лучше, сколько кластеров должно быть и есть ли соответствие между полученными группами и реальными объектами? Это вопросы кластерной достоверности. Для ответа на них разрабатываются специальные методы, вводятся критерии и меры качества. Обзор методов дан в работе [1], где показано, что выбор как критериев классификации, так и критериев качества распределений векторов по кластерам зависит от природы данных и задачи исследования.

Предлагаемая работа касается классификации многоспектральных изображений поверхности Земли, полученных дистанционными методами. Специфика состоит в большом объеме и разнообразии данных, а также часто почти в полном отсутствии априорной информации. Кластерные алгоритмы, применяемые в дистанционном зондировании, делятся в основном на три типа: по k-центрам, иерархические и основанные на многомерной гистограмме [2]. Алгоритмы, основанные на гистограмме, имеют ряд преимуществ в сравнении с другими: не нужно задавать число кластеров, как для ал-

* Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (проект № 05-07-90057).

горитмов классификации по *k*-центрам, и они существенно быстрее иерархических. Недостаток их в том, что многомерная гистограмма требует много оперативной памяти. Однако хранение лишь присутствующих на изображении векторов позволяет работать с данными большой размерности, не прибегая к использованию дисковой памяти. Различные системы хеширования для доступа к данным при формировании гистограммы предложены в [3, 4]. В исследовании использован алгоритм разделения многомерной гистограммы по унимодальным кластерам, соответствующим локальным максимумам гистограммы [3]. Алгоритм находит все локальные максимумы гистограммы, которых слишком много. Они могут соответствовать очень близким кластерам или оказаться мелкими случайными всплесками на поверхности гистограммы. В некоторых случаях кластеры можно объединить. Однако для задач распознавания важно иметь унимодальные кластеры. При объединении кластеров свойство унимодальности может быть потеряно. Изменить детальность распределения можно до классификации, объединив векторы уменьшением числа уровней квантования векторного пространства.

Предлагается: гистограммным методом классификации получить ряд распределений векторов для различного числа уровней квантования векторного пространства, ввести критерий и меру качества распределения, найти лучшие распределения построенного ряда в смысле предложенной меры. Быстрота гистограммного алгоритма позволяет многократно использовать его при построении ряда.

Квантование. Число уровней квантования N равно числу возможных целых значений, принимаемых многоспектральным вектором по каждому спектральному каналу. Начальное число уровней квантования $N_0 = 256$, $N < N_0$. Размер ячейки для произвольного уровня квантования положим вещественным: $kf = (N_0 - 1)/(N - 1)$. L – число спектральных каналов, $f = [f(1), f(2), \ldots, f(L)]$ – многоспектральный вектор изображения, а $g = [g(1), g(2), \ldots, g(L)]$ – вектор, в который преобразуется f в результате квантования:

$$g(k) = \text{entier}\left(\frac{f(k)}{kf}\right), \quad k = 1, \ldots, L.$$

Для дальнейшей классификации используются новые векторы g. Однако сохраняется связь новой системы векторов со старой: значение гистограммы для нового вектора получается суммированием значений гистограммы старых соответствующих векторов.

Алгоритм классификации. Детальное описание конкретной реализации представлено в [4]. Суть алгоритма состоит в нахождении локальных максимумов многомерной гистограммы в дискретном векторном пространстве и отнесении векторов к соответствующим максимумам. Для каждого вектора строится элементарный граф по направлению максимума положительного градиента плотности вероятности в списке соседей. Векторы связываются в «деревья» с помощью элементарных графов. Когда граф достигает локального максимума, вся цепочка векторов относится к тому же кластеру, что и максимум (корень дерева). Границы кластеров соответствуют долинам гистограммы. Векторы на границе кластеров классифицируются по тому же принципу максимума положительного градиента. Если гистограмма содержит плато, то предусмотрены меры от зацикливания. Классификация жесткая: каждый вектор относится только к одному кластеру. Трассирование

элементарных графов обеспечивает линейную зависимость количества операций от числа векторов. Важно отметить, что при построении графов вычисления производятся только со скалярными значениями гистограммы. Операции с многомерными векторами осуществляются только на этапе построения списка ближайших соседей каждого вектора. Векторы упорядочиваются по возрастанию, и поиск соседей также является быстрой процедурой. В результате классификации векторы распределяются по связным унимодальным кластерам.

Критерий и мера качества. Общее требование к качеству классификации – это хорошее разделение кластеров в векторном пространстве. Требование выполняется, когда векторы в кластерах находятся далеко от границ [5]. Если векторов на границах кластеров нет, значит, они полностью разделены в пространстве признаков. Но для изображений поверхности Земли, как показывают измерения, такая ситуация – редкость. Часто кластеры тесно расположены хотя бы в нескольких направлениях пространства. Присутствует обычно ряд кластеров для какого-либо природного объекта, например: лес может состоять из разных пород, вода иметь различную глубину, снег находиться в разных фазах таяния, облака иметь разную прозрачность и т. д. Можно предположить, что кластеры объектов одной природы расположены близко, на их границах много векторов, т. е. значение гистограммы на границах велико. Напротив, различные природные объекты разделены лучше, и на их границах значение гистограммы мало. Определим критерий качества классификации: чем больше кластеров имеет низкое значение гистограммы на границе, тем лучше классификация. Введем меру качества сначала для отдельного унимодального кластера (число уровней квантования N):

$$M^j(N) = \frac{\dfrac{1}{B^j(N)} \sum\limits_{i=1}^{B^j(N)} h_i^j(N)}{H^j(N)}, \qquad (1)$$

где $h_i^j(N)$ – значение гистограммы в i-й точке границы кластера j; $B^j(N)$ – число точек границы кластера; $H^j(N)$ – максимальное значение гистограммы.

Мера (1) определяется как отношение среднего значения гистограммы на границе к максимальному значению гистограммы кластера. В (1) всегда $M^j(N) \le 1$, так как кластеры унимодальны. Чем меньше значение меры (1), тем лучше кластер. Хорошие кластеры могут быть острыми при быстром убывании гистограммы от точки максимума. В этом случае кластеры компактны, т. е. большая часть векторов находится у точки максимума. Если же гистограмма убывает медленно, то небольшие значения (1) могут быть получены для протяженного кластера, т. е. большая часть векторов кластера удалена от других кластеров. Таким образом, выполняется требование, которое предъявляется к качеству классификации в задачах кластерной достоверности [6]: компактность кластеров и их удаленность друг от друга. Меру качества распределения в целом определим как среднее значение по кластерам, число которых $K(N)$:

$$M(N) = \frac{1}{K(N)} \sum_{j=1}^{K(N)} M^j(N). \qquad (2)$$

Небольшие значения (2) соответствуют тому, что векторов на границах мало, т. е. качество распределения хорошее. При определении границ кластеров используется список ближайших соседей векторов в пространстве признаков, который был построен ранее для алгоритма классификации. Имея этот список, легко узнать граничные векторы всех кластеров за один его просмотр после классификации векторов. Вычисления меры (1) производятся со скалярными значениями гистограммы. Если значение меры (2) становится равным нулю, то меру уже нельзя использовать для дальнейшего улучшения качества, т. е. для поиска таких распределений, в которых кластеры еще дальше удалены друг от друга. В этом случае можно предложить другие меры [1]. Однако в задачах дистанционного зондирования это маловероятно. К тому же при тесном контакте кластеров меры, основанные на замене компактности дисперсией [5], порождают неопределенность, потому что дисперсия кластера зависит не только от остроты пика гистограммы, но и от диаметра кластера, а диаметр определяет расстояние до ближайших кластеров.

Эксперименты. Меняя число уровней квантования N, получим ряд распределений векторов и выберем лучшие распределения по минимальным значениям меры (2). Проиллюстрируем выбор лучшей классификации для двухспектрального изображения. На диаграмме плотности распределения двумерных векторов (рис. 1) показаны кластеры, построенные алгоритмом для различного числа уровней квантования. Обрабатывался фрагмент изображения Западной Сибири, полученного 17 апреля 2003 года со спутника NOAA17.

По оси абсцисс отображена яркость в ближнем инфракрасном диапазоне (R), по оси ординат – в зеленой части спектра (G). На диаграмме можно увидеть, что в целом векторы образуют два больших кластера и один небольшой хорошо отделенный. Для пяти уровней квантования значение меры $M(5) = 0{,}27$ (рис. 1, *a*). Алгоритм построил три кластера: кластер 1 соответствует заснеженной поверхности, кластер 2 – оттаявшей, кластер 3 – воде озер. Рис. 1, *b* соответствует лучшей классификации: мера изолированности (2) достигает минимума $M(8) = 0{,}11$ для восьми уровней квантования ($N = 8$). Хотя здесь также получено три кластера, но граница между кластерами проходит точнее, чем на рис. 1, *a*, по менее плотной области. Увеличение детальности привело к расщеплению кластера 1 на два, граница между новыми кластерами лежит в довольно плотной области диаграммы (рис. 1, *c*), поэтому для этой классификации значение меры (2) стало больше: $M(10) = 0{,}14$.

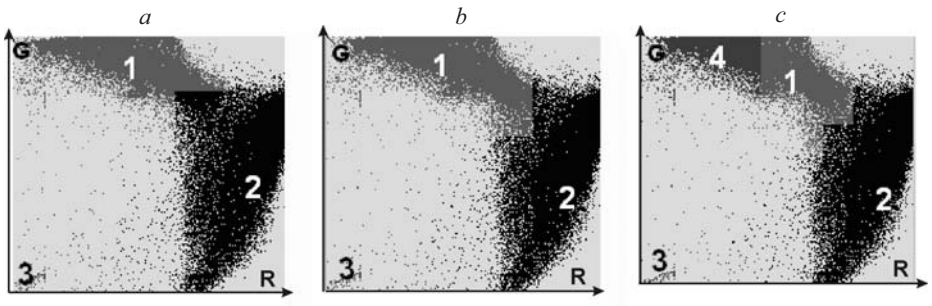

Рис. 1. Двумерные векторные распределения для различного числа уровней квантования

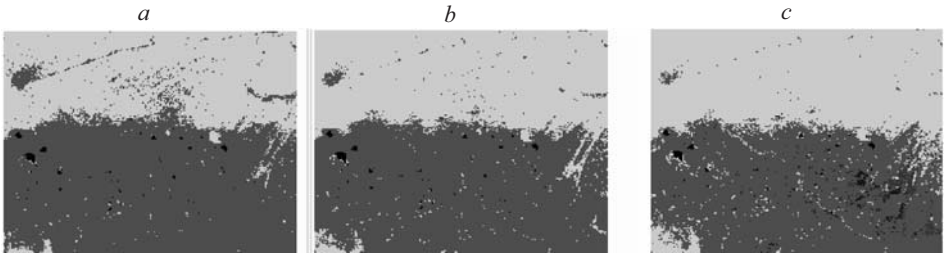

Рис. 2. Кластерные карты ряда классификаций фрагмента изображения Западной Сибири: вверху находится заснеженная поверхность, внизу – оттаявшая, черные пятна – озера. Темное пятно слева вверху – г. Омск, косые светлые полосы справа – Сузунские ленточные боры

Кластерные карты двухканального изображения, соответствующие диаграммам на рис. 1, *a*, *b*, представлены на рис. 2, *a*, *b*. Часть заснеженной поверхности из-за ошибки квантования отнесена (слишком грубо) к оттаявшей, как следует из рис. 2, *a*. Лучшая классификация дана на рис. 2, *b*.

При размерности данных больше двух трудно найти наглядную интерпретацию распределений, поэтому актуальна оценка меры качества при определении структуры таких данных. Для пяти спектральных каналов (два в видимой части спектра, три в различных частях инфракрасного диапазона) лучшая классификация того же фрагмента представлена на рис. 2, *c*. Значение меры очень мало: $M(8) = 0{,}08$, число кластеров восемь, и они хорошо изолированы. Сузунские боры представлены несколькими кластерами. Однако это не единственный минимум меры для данного изображения. Разнообразие небольших объектов оттаявшей поверхности порождает большое число хорошо разделенных кластеров для высокого числа уровней квантования. Следующий по возрастанию минимум меры имеет все еще небольшое значение ($M(25) = 0{,}14$) и соответствует 270 кластерам. После фильтрации

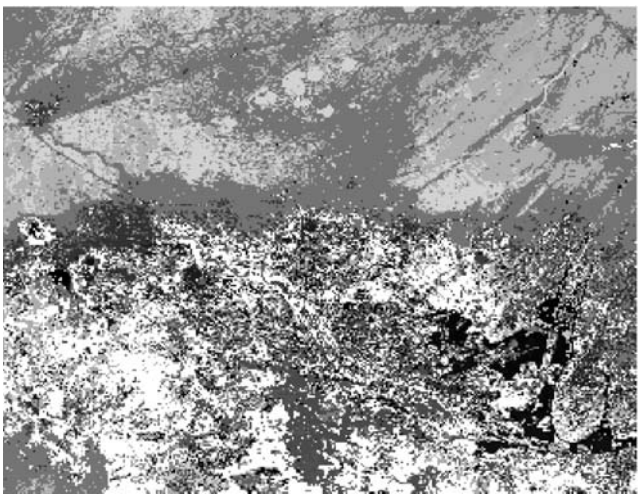

Рис. 3. Кластерная карта для одной из лучших классификаций фрагмента изображения Западной Сибири. Видны реки, железные дороги (Иртыш расположен по диагонали от г. Омска). Снег представлен небольшим числом кластеров. Светлое пятно вверху посередине – озеро Чаны под снегом

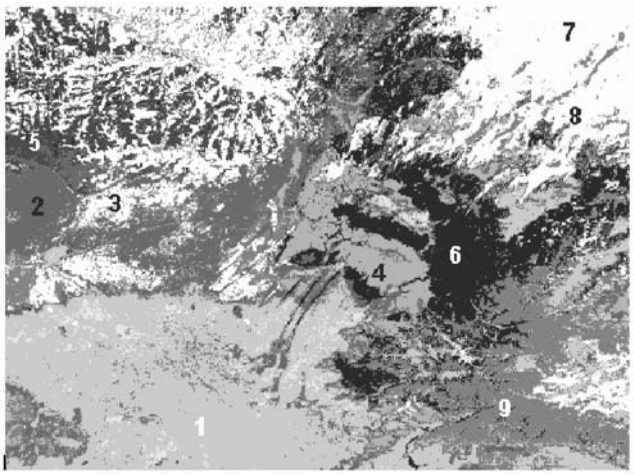

Рис. 4. Кластерная карта лучшей классификации изображения Западной Сибири (17 апреля 2003 г.). Слева расположен г. Омск, косые полосы посередине – ленточные боры. Наиболее крупные кластеры: 1 – оттаявшая поверхность, 2–4 – различные фазы таяния снега, 5, 6 – хвойный лес под тающим снегом, 7 – густое облако, 8, 9 – полупрозрачные облака

совсем мелких кластеров оставлено 74 кластера. Кластерная карта для 25 уровней квантования (увеличенный фрагмент) показана на рис. 3.

Для заснеженных территорий, однородных по природе, распределения менее расчленены. Кластерная карта лучшего распределения для полного пятиспектрального изображения Западной Сибири (8,3 Мбайт) представлена на рис. 4. Здесь число уровней квантования $N = 12$, значение меры качества кластера $M = 0{,}1$. Получено 7888 различных векторов и 16 кластеров. На карте указано 9 наиболее крупных кластеров. При дальнейшем увеличении числа уровней квантования значение меры существенно возрастает.

Представим классификацию облаков. Облака обычно однородные объекты по спектральным признакам. На рис. 5 показана кластерная карта одно-

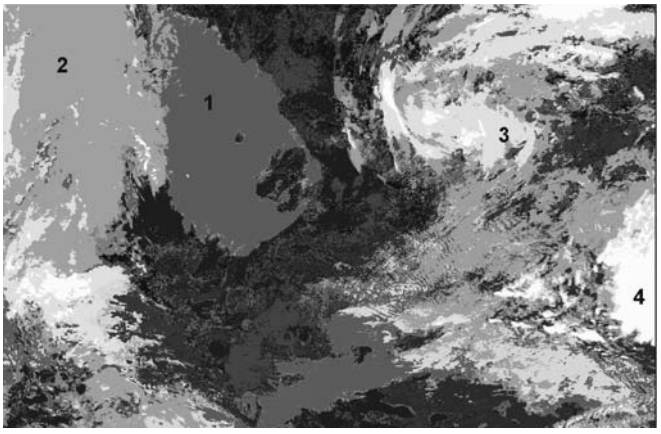

Рис. 5. Кластерная карта лучшей классификации изображения Сибири (26 марта 2003 г.). Светлыми оттенками выделены облака: 1 – низкие кучевые, 2 – высокие перистые, 3, 4 – еще более высокие; в правой части карты – циклон

го из лучших распределений векторов для пятиспектрального снимка территории Сибири, полученного со спутника NOAA17. Минимум меры $M(14) = 0{,}11$, получено 29 кластеров, большинство из них относится к поверхности Земли. Облакам соответствует несколько кластеров.

Заключение. Рассмотрена кластеризация многоспектральных изображений в два этапа: объединение векторов с помощью уменьшения числа уровней квантования и классификация по унимодальным кластерам многомерной гистограммы. Параметр – число уровней квантования – определяет соотношение во взаимодействии этих двух способов группирования данных. Предложенная мера качества классификации является индикатором разделения кластеров. Минимизируя меру как функцию числа уровней квантования, можно получить лучшие распределения по мере в смысле разделения кластеров. Эксперименты показывают, что унимодальные кластеры лучших распределений соответствуют представительным информационным классам зондируемой поверхности Земли.

СПИСОК ЛИТЕРАТУРЫ

1. **Halkidi M., Batistakis Y., Vazirgiannis M.** On clustering validation techniques // Journ. of Intelligent Information Systems. 2001. **17**, N 2–3. P. 107.
2. **Gong P., Howarth P. J.** An assessment of some factors influencing multispectral land-cover classification // Photogrametric Eng. and Remote Sensing. 1990. **56**, N 5. P. 597.
3. **Narendra P. M., Goldberg M.** A non-parametric clustering scheme for LANDSAT // Pattern Recogn. 1977. **9**. P. 207.
4. **Sidorova V. S.** Separating of the multivariate histogram on the unimodal clusters // Proc. of the IASTED Intern. Conf. "Automation, Control, and Information Technology (ACIT'2005)". Anaheim – Calgary – Zurich: ACTA Press, 2005. P. 50.
5. **Fukunaga K.** Introduction to Statistical Pattern Recognition. New York – London: Academic Press, 1972.
6. **Davies D. L., Bouldin D. W.** A cluster separation measure // IEEE Trans. Pattern Anal. Machine Intel. 1979. **1**, N 4. P. 224.

Поступила в редакцию 14 июля 2006 г.