Квантование векторного пространства и классификация многомерной гистограммы [1]

Валерия Сидорова
Институт Вычислительной Математики и Математической Геофизики СО РАН,
Новосибирск, Россия
svs@ooi.sscc.ru

Аннотация

Рассматривается задача уменьшения числа кластеров многомерной гистограммы с помощью квантования векторного пространства. Предлагаются критерий и мера качества классификации для выбора лучшей классификации. Алгоритм применяется для многоспектральных спутниковых изображений поверхности Земли, и результаты иллюстрируются.

Ключевые слова: *дистанционное зондирование, обработка изображений, квантование, кластерный анализ, многомерная гистограмма.*

1. ВВЕДЕНИЕ

В дистанционном зондировании для анализа многоспектральных данных используются кластерные методы: распределение по к-центрам, иерархические, по многомерной гистограмме [1]. В последние годы появились новые алгоритмы, например [2-4,6]. В виду огромного числа пикселей на изображениях и большого числа спектральных каналов, многие методы затруднительно применять ввиду значительных вычислительных затрат. Кроме того, в предлагаемых методах требуется задание числа кластеров (разделительные методы по к-центрам) или других параметров, которые для большого объема разнообразных данных трудно определить. Методы, основанные на многомерной гистограмме, лишены этих недостатков. Правда, для ее хранения требуется много памяти. Однако хранение только присутствующих на изображении векторов и гистограммы в виде упорядоченного списка существенно сокращает объемы выделяемой оперативной памяти. Доступ к списку осуществляется различными системами хеширования, одна из которых была предложена авторами[7].

Мы используем алгоритм кластеризации, известный давно[5]. Он разделяет многомерную гистограмму по унимодальным кластерам. Алгоритм не использует никаких предположений о функциях распределения различных классов, не требует никаких априорных данных (число кластеров, количество итераций и т.д.). Он является не итеративным и быстрым, т.к. гистограмма строится за один просмотр изображения, время классификации линейно зависит от числа векторов.

Хотя распределение векторов по кластерам позволяет существенно сократить объемы информации для анализа, все же и кластеров получается много. В этом исследовании мы, во-первых, решаем задачу сокращения числа кластеров. Полученные кластеры могут оказаться очень близко друг к другу. Их можно объединить, используя другие методы, например, самый простой - по порогу близости значений векторов в точках максимумов гистограммы. Однако пороговые и многие другие кластерные методы не гарантируют получение унимодальных кластеров, а свойство кластеров иметь один максимум является важным для дальнейшего анализа, для возможности применения классических методов распознавания. Применить этот же алгоритм к полученным центрам кластеров мы не можем из-за его специфики. Он хорошо работает на больших объемах плотных данных, чтобы гистограмма могла аппроксимировать плотность вероятности многомерных векторов, чтобы у векторов могло быть достаточно много ближайших соседей. Поэтому используется другой подход для сокращения числа кластеров – предварительное уменьшение числа различных многоспектральных векторов.

Ранее предлагалось [7] округлять значения компонент векторов, убирая младшие биты, или сглаживать гистограмму по соседям каждого вектора. Но такое сглаживание не может уменьшить число кластеров, если образуются группы векторов, не имеющие соседей. С другой стороны, срезание одного бита в каждом спектральном канале эквивалентно уменьшению числа уровней квантования сразу вдвое. Таким образом, размер ячейки квантования увеличивается, и в нее может попасть несколько соседей, т.е. происходит предварительное объединение векторов пороговым методом. Число новых векторов и число кластеров уменьшается скачком. Чтобы управлять детальностью кластеризации плавно, предлагается число уровней квантования менять постепенно. В новой системе векторов кластеры могут быть дальше друг от друга, но не ясно, насколько они изолированы. Мы получили совокупность распределений для различного числа уровней квантования пространства векторов, как для параметра.

Для этой совокупности распределений решим еще одну задачу: найдем лучшее (или лучшие) в смысле изолированности кластеров. Лучшему распределению будут соответствовать определенное число кластеров и уровень квантования. Такой подход к решению задачи классификации предложен давно [8], но мало использовался, т.к. многократная классификация данных требует больших вычислительных затрат. В последние годы вопросы качества классификации, выбора мер и критериев для различных кластерных схем и алгоритмов активно исследуются [9-12]. В этой работе будет предложена дешевая мера качества классификации для алгоритма разделения гистограммы по унимодальным кластерам.

[1] Работа выполнена частично при финансовой поддержке Российского фонда фундаментальных исследований (проект № 05-07-90057).

2. АЛГОРИТМ КЛАССИФИКАЦИИ

Кратко опишем алгоритм классификации. 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. Если все соседи вектора имеют плотность меньшую, чем он сам, это значит, что гистограмме в этой точке векторного пространства соответствует локальный максимум и данный вектор является корнем соответствующего ориентированного графа. Вектора связываются в деревья с помощью элементарных графов. Когда граф достигает локального максимума, то вся цепочка векторов относится к тому же кластеру, что и максимум (корень). Таким образом, формируются унимодальные кластеры. Вектора на границе кластеров относятся к кластерам по тому же принципу максимума положительного градиента. Если гистограмма содержит плато, то предусмотрены меры от зацикливания. В результате работы алгоритма гистограмма разделяется на непересекающиеся области. Пространство векторов разделяется на соответствующие непересекающиеся кластеры (жесткая классификация). Граница кластера соответствует долинам гистограммы. Важно отметить, что при построении графов операции производятся только со скалярными значениями гистограммы. Трассирование графов обеспечивает линейную зависимость количества операций от числа векторов. Операции с многомерными векторами осуществляются только на этапе построения списка ближайших соседей каждого вектора. Однако вектора хранятся в виде упорядоченного по возрастанию списка, поэтому поиск соседей также является быстрой процедурой. Более подробно в [11-13]. Благодаря быстроте алгоритма возможно многократное его применение для различных уровней квантования. Также следует учесть, что, если двигаться от большего числа уровней квантования, то возникает экономия при пересчете гистограммы, соседей векторов, а также происходит существенное уменьшение числа различных векторов.

3. КВАНТОВАНИЕ

Квантование позволяет предварительно объединить вектора и может быть использовано для уменьшения их числа. Число уровней квантования *N* равно числу возможных целых значений, принимаемых многоспектральным вектором по каждому спектральному каналу. Начальное число уровней квантования *N0*=256. Размер ячейки для произвольного уровня квантования положим вещественным *kf=(N0-1)/(N-1)*. Пусть *L* число спектральных каналов, ***f***=[*f(1),f(2),…,f(L)*] – многоспектральный вектор изображения, а ***g***=[*g(1)g(2),…g(L)*] - вектор, в который преобразуется ***f*** в результате квантования:

$$g_k = entier\left(\frac{f(k)}{kf}\right), k = 1,...,L.$$

Если *kf* >1, то число различных векторов ***g*** в новой системе может быть меньше числа векторов ***f*** в старой системе.

4. МЕРА КАЧЕСТВА КЛАССИФИКАЦИИ

Определение хорошей классификации – хорошей изолированности кластеров, дано в [9] и широко используется: кластеры должны быть компактны и далеко отстоять друг от друга. Под компактностью понимается, что большая часть векторов кластера сосредоточена в центре, подальше от границ. Часто это дисперсия. Например, мера качества классификации ***DB*** по ***nc*** кластерам для одномерных векторов [10]:

$$R_{ij} = (s_i + s_j)/d_{ij},$$

$$DB_{nc} = \frac{1}{nc}\sum_{i=1}^{nc} R_i$$

$$R_i = \max_{i=1,...,nc, i\neq j} R_{ij}, i = 1,...,nc$$

здесь ***i, j*** – номера кластеров, ***s*** – дисперсия кластера, ***d*** – расстояние между кластерами. Лучшая классификация соответствует минимуму меры ***DB.*** Предлагаемые меры работают в пространстве векторов, поэтому их вычисление требует значительных затрат, особенно, если пространство многомерное.

Используемый нами алгоритм работает со значениями гистограммы, попытаемся и меру качества распределения по кластерам выразить через скалярные значения гистограммы. Сначала введем меру качества отдельного унимодального кластера (число уровней квантования *N)*:

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

где $h_i^j(N)$ значение гистограммы в *i*-той точке границы кластера *j*, $B^j(N)$ число точек границы кластера, $H^j(N)$ максимальное значение гистограммы.

Мера (1) определяется как отношение средней высоты гистограммы на границе к максимальному значению гистограммы кластера. Под границей кластера понимаются все его граничные точки, их номера в списке векторов легко определяются для данного алгоритма. $M^j(N)$ <=1. Чем меньше значение меры (1), тем лучше кластер. Для функции, имеющей один максимум, и убывающей от точки максимума, маленькое значение меры (1) означает: 1) или значение гистограммы быстро убывает при удалении от точки максимума, и кластер имеет маленькую дисперсию и компактен, 2) или расстояние от точки максимума до границы кластера велико, и значит, велико расстояние до соседних кластеров. В обоих случаях, доля граничных точек кластера существенно меньше доли внутренних, и, поэтому кластер хорошо отделен от соседей. Меру качества распределения всех векторов определим как среднее значение по кластерам, число которых *K(N)*:

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

Теперь, меняя число уровней квантования N, получим последовательность классификаций векторов и выберем лучшие распределения по минимальным значениям меры (2).

5. ЭКСПЕРИМЕНТЫ

Рассмотрим фрагмент трехспектрального спутникового изображения тающих снегов поверхности Земли (два канала в видимой части спектра, третий в ближнем инфракрасном), рис.1 вверху. Исследовалось поведение меры качества классификации (2) в соответствии с увеличением числа уровней квантования пространства векторов. Наименьшее значение меры качества (2) $M(8)$=0.11 соответствовало восьми уровням ($N=8$), при этом получилось 50 различных векторов и три кластера. Кластеры на карте соответствовали трем основным природным объектам, хорошо разделенным в пространстве признаков (рис.1 слева). Следующий по возрастанию минимум меры (2) соответствует 10 уровням квантования и равен $M(10)=0.18$. Число различных векторов при этом равно 204, а число кластеров равно 9. Новые кластеры появились в основном в небольшом участке оттаявшей поверхности. Снежный покров разделился на два кластера (рис.1 справа). Более детальный уровень классификации приводит к значительному увеличению меры $M(N)$, т.е. к ухудшению качества. Хотя для данного изображения увеличение числа уровней квантования улучшает детальность классификации, но качество классификации – разделение кластеров, ухудшается. Снежный покров довольно однородный объект, и спектральные характеристики различных участков близки друг другу.

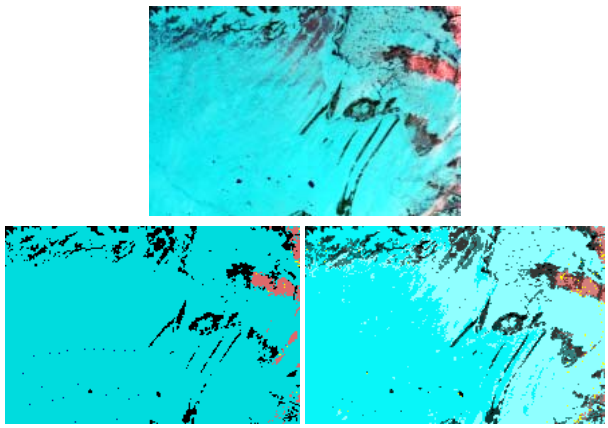

Рис.1. Вверху снимок тающих снегов: светлый тон - снежный покров, темный – хвойный лес, оконтуренный участок - оттаявшая поверхность. Внизу лучшие классификации.

Иное поведение меры наблюдалось для фрагмента другого пятиспектрального спутникового изображения поверхности Земли над территорией Западной Сибири, полученного в апреле со спутника NOAA. На рис.2 слева представлено изображение для трех первых каналов в цветосовмещенном формате (RGB-файл): два канала в видимой части спектра, а третий в ближнем инфракрасном диапазоне. На рис.2 справа изображение в пятом инфракрасном канале на длине волны 12 мкм. Абсолютный минимум меры соответствует девяти уровням квантования $M(8)$=0.05, здесь получилось три кластера: озера, поверхность Земли под снегом и оттаявшая поверхность (рис.2 слева). Спектральные характеристики этих объектов практически не пересекаются для данного уровня детализации (здесь 1894 различных вектора).

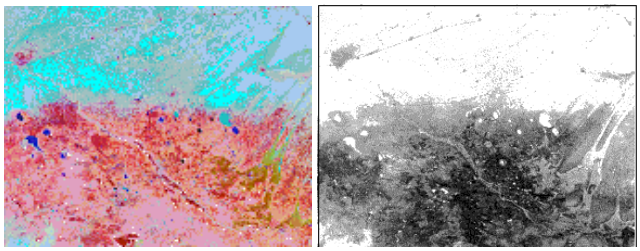

Рис.2. Фрагмент изображения Западной Сибири. В верхней части тающие снега, в нижней - оттаявшая поверхность Земли. Пятно слева вверху – г. Омск., слева - Ленточные боры.

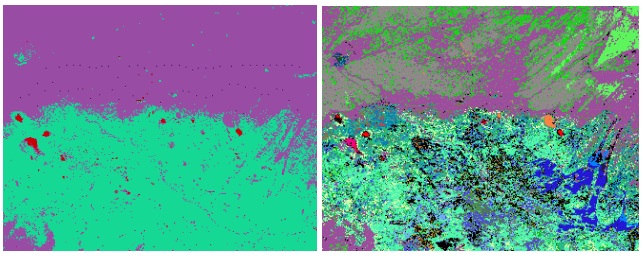

Рис.3. Лучшие классификации.

На рис.3 справа показана следующая по возрастанию минимума классификация $M(26)$=0.13. Здесь получено очень много кластеров, но после фильтрации очень маленьких (1, 2 пикселя) осталось 36 унимодальных кластера (19447 различных векторов). Большое число кластеров вполне соответствует тому, что оттаявшая поверхность представляет собой чрезвычайно пеструю картину во всех спектральных каналах. Однако маленькие объекты хорошо различимы, что подтверждает маленькое значение меры для довольно большой детальности классификации.

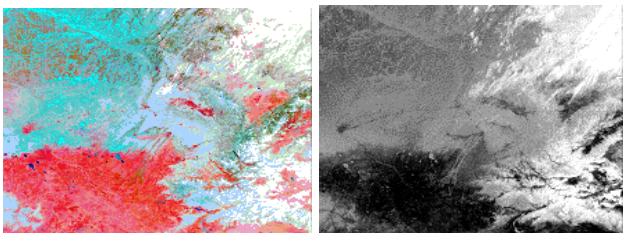

Рис.4. Слева изображение трех первых спектральных каналов в цветосовмещенном формате на спутниковом снимке Западной Сибири в апреле, справа - пятый спектральный канал.

Приведем также результаты одной из лучших классификаций для большого (8.3 МБ) пяти-спектрального изображения (Рис.4 исходные снимки). Лучшая в смысле меры (2) классификация соответствует делению на два кластера: оттаявшая поверхность и та, что под снегом. Но мы покажем следующую по величине меры на рис.5; здесь уровень квантования N=17, значение меры качества кластера M=0.162. Получено 18838 различных векторов и 32 значимых кластера. На карте указаны девять наиболее крупных кластеров.

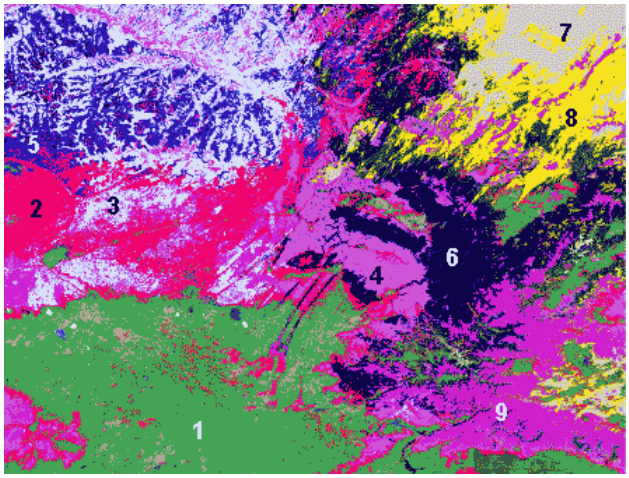

Рис.11. Лучшая классификация. Наиболее крупные кластеры: 1 – оттаявшая поверхность, 2,3,4-различные фазы таяния снега, 5,6 – тающий снег над хвойным лесом, 7 – густое облако, 8 и 9 полупрозрачные облака.

6. ЗАКЛЮЧЕНИЕ

Предложена кластеризация многоспектральных изображений в два этапа. Предварительное объединение векторов с помощью увеличения ячейки квантования (уменьшения числа уровней), и затем классификация новой системы векторов по унимодальным кластерам с помощью многомерной гистограммы. Параметр – число уровней квантования – определяет соотношение во взаимодействии этих двух способов группирования данных. Предложенные критерий и мера качества классификации являются индикаторами разделения кластеров. Минимизируя меру как функцию уровня квантования, можно получить лучшие распределения. Анализ рассмотренных многоспектральных спутниковых изображений показал, что такие распределения соответствовали различным значениям уровня в связи с различной структурой изображений, но этот уровень оказался существенно меньше 256. Таким образом, вместо объединения кластеров для уменьшения их числа, мы получили сравнительно небольшое число унимодальных кластеров, в среднем хорошо разделенных в пространстве признаков.

7. ССЫЛКИ

[1] Gong P. & P.J. Howarth. An assessment of some factors influencing multispectral land-cover classification. Photogrametric Engineering and Remote Sensing, 56(5), 1990, 597-603.

[2] Wang, W., Yang, J., Muntz, R. STING: A statistical information grid approach to spatial data mining. Proceedings of 23rd VLDB Conference, 1997.

[3] Ester, M., Kriegel, H-P. Sander, J., Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of 2nd Int. Conf. On Knowledge Discovery and Data, 1996, 226-231.

[4] Ng, R., Han, J. Effecient and Effictive Clustering Methods for Spatial Data Mining. 20th VLDB Conference, Santiago, Chile, 1994.

[5] P.M.Narendra and M.Goldberg, A non-parametric clustering scheme for LANDSAT. Pattern Recognition, 9, 1977, 207-215.

[6] Sheikholeslami G., Chatterjee S. and Zhang A. Wavecluster: a multi-resolution clustering approach for very large spatial databases. Proceedings of 24th Conf. on Very Large Data Bases, 1998.

[7] Сидорова В.С. Многомерная гистограмма и разделение векторного пространства признаков по унимодальным кластерам. Труды конференции GraphiCon'2005. 2005, 267-274.

[8] Keinosuke Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York and London, 1972.

[9] D. L. Davies and D. W. Bouldin. A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intel., vol. 1, no. 4, 1979, 224–227

[10] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.*, vol. 3, no. 3, 1973, 32–57.

[11] Rezaee, R, Lelieveldt, B.P.F., Reiber, J.H.C (1998). A new cluster validity index for the fuzzy c-mean, Pattern Recognition Letters, 19, 1998, 237-246.

[12] Xie, X. L, Beni, G. A Validity measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and machine Intelligence, Vol.13, No4, 1991.

Об авторе

Валерия Сидорова – научный сотрудник Института Вычислительной Математики и Математической Геофизики СО РАН, лаборатории Обработки изображений.

Адрес: Новосибирск, 630090, Россия, просп. Лаврентьева 6, ИВМиМГ.

Телефон. (383) 307-23-32,

svs@ooi.sscc.ru

Quantizing Vector Space and Clustering Multivariate Histogram

Abstract

Quantizing vector space for reducing the cluster number of the multidimensional histogram is proposed. The selection criterion for the number of quantization levels is offered, the quality measure of clustering too. Satellite images of the Earth surface were researched by means of the technique.

Keywords: Remote sensing, Image processing, quantization, clustering, multidimensional histogram.

About the author

Valeria Sidorova is a scientific researcher at Institute of Computational Mathematics and Mathematical Geophysics, Department of Image Processing. Her contact email is svs@ooi.sscc.ru