ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

Кластеризация гиперспектральных изображений поверхности земли[1]

Сидорова Валерия Сергеевна, научный сотрудник
Институт вычислительной математики и математической геофизики
Сибирского отделения Российской академии наук

Предложен алгоритм кластеризации гиперспектральных изображений Земли. Алгоритм состоит из двух этапов. На первом этапе число спектральных каналов сокращается. На втором этапе производится кластеризация по методу, выделяющему кластеры по разделимости. Приводится пример.

Ключевые слова: *гиперспектральные данные, дистанционное зондирование, кластеризация, многомерная гистограмма, кластерная разделимость.*

Хотя гиперспектральная съемка предполагает 220 каналов в каждой точке изображения поверхности, однако совсем не обязательно, что все эти каналы равноценны. Для поверхности Земли может быть существенно меньше каналов. Изображение может быть вытянуто в нескольких направлениях, в остальных могут быть случайные точки. Поэтому можно сократить число каналов. Различие объектов может быть связано с различной степенью яркости в различных каналах.

Здесь будет рассмотрено известное изображение территории Северной Америки размером 145x145, содержащее 220 каналов. Сначала выполняется квантование пространства признаков[1] методом Якоби. Правило квантования, обеспечивающее наименьшую потерю информации, требует различного подхода в различных направлениях, а именно: квантование должно сохранять ячейку квантования в форме гиперкуба (а не гиперпараллелепипеда). Это условие будет выполнено, если число уровней квантования вдоль каждой оси собственного пространства пропорционально квадратному корню из соответствующего собственного числа. (Собственное число характеризует разброс вдоль оси), а именно:

$$\frac{N_{e1}}{\lambda_{e1}} = \frac{N_{e2}}{\lambda_{e2}} = \ldots = \frac{N_{ek}}{\lambda_{ek}} \quad ,(1)$$

где , N_{e2} , … , N_{ek} числа уровней квантования вдоль для соответствующих собственных векторов по k ортонормированным осям, а λ^2_{e1} , λ^2_{e2} , … , λ^2_{ek} собственные числа в квадрате.

Зададим максимальное число уровней квантования в собственном пространстве равным N_{em} . Тогда, в соответствии с пропорциями (1) может быть найдено число уровней квантования и по другим осям собственного пространства. Для задач кластеризации это число должно быть больше или равно 2, иначе эта компонента одинакова для всех векторов и никакой роли в кластеризации не играет. Расположим собственные числа по убыванию: 5055,5, 2947,7, 1246,3, 781 . Максимальное значение для N_{em} , двубайтового целого числа равно. Из (1) видно, что трех каналов достаточно. Вообще этот метод кластеризации мультиспектральных изображений допускает до 10 каналов, все зависит от мощности компьютера. (Дальнейшее уменьшение собственных чисел делает число уровней квантования меньшим единицы). Таким образом, мы получаем сокращение размерности пространства признаков. В этих трех каналах изображение будет выглядеть как на рис. 1.

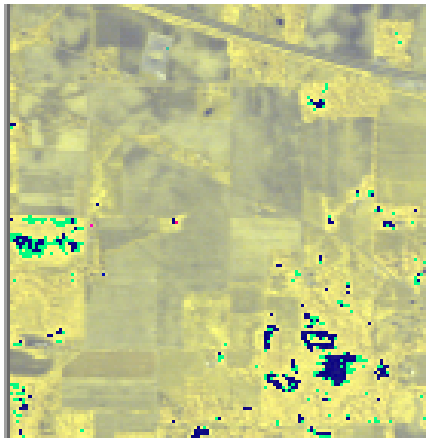

Рис.1

Теперь применим кластеризацию. Мы выбрали иерархическую[основанную на выборе разделимости кластеров[2]. Чем меньше разделимость, тем лучше кластеры отделены друг от друга. Достоинство его в том, что он находит все унимодальные кластеры, но детальность данных в каждом из них определяется своя. Увеличивая детальность данных на каждом иерархическом уровне, получаем больше кластеров. Но в каждом кластере их столько, сколько нужно, чтобы обеспечить данную разделимость подкластеров. И в целом их получается гораздо меньше, чем тогда, когда детальность

[1] Работа выполнена частично при финансовой поддержке Российского фонда фундаментальных исследований (проект № 16-07-00066).

ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

выбирается одна и та же для всего изображения. Таким образом, мы получим небольшое число хорошо отделенных кластеров. Более подробно.

Основной алгоритм использует параметр детальности - число уровней квантования n векторного пространства. Пусть их начальное число $n<n0$, $n0=256$. Размер ячейки для произвольного уровня квантования $kf=(n0-1)/(n-1)$. Пусть L – число признаков, $f=[f(1),f(2),...,f(L)]$ – вектор признаков, $g=[g(1),g(2),...,g(L)]$ – вектор, в который преобразуется f:

$$g(k)=\left[\frac{f(k)}{kf}\right],\ k=1,...,L,$$

где $[\cdot]$ целая часть числа.

Основной алгоритм отыскивает такое число n (и соответствующие g), при котором полученное методом Нарендры[3] распределение векторов по унимодальным кластерам дает минимум мере разделимости кластеров, в диапазоне изменения n.

$$m^j(n)=\frac{1}{B^j(n)*H^j(n)}\sum_{i=1}^{B^j(n)}h_i^j(n),\ (1)$$

$$m(n)=\frac{1}{K(n)}\sum_{j=1}^{K(n)}m^j(n),\ (2)$$

где $h_i^j(n)$ значение гистограммы в i-той точке границы кластера j, $B^j(n)$ число

точек границы кластера j, $H^j(n)$ максимальное значение гистограммы кластера j.

Алгоритм Нарендры является жестким, т.е. каждый вектор принадлежит только одному кластеру. С помощью графов он разделяет векторное пространство признаков по унимодальным кластерам, модальные векторы которых соответствуют локальным максимумам гистограммы. Направление элементарного графа в каждой точке дискретного векторного пространства соответствует градиенту гистограммы и определяется по окрестности ближайших соседей точки, т.е. векторов, каждая компонента которых отстоит от этой точки не далее, чем на единицу. Список соседей каждого вектора может быть вычислен заранее. В результате классификации границы кластеров проходят по долинам гистограммы, то есть по областям низкой плотности векторов. В памяти хранятся только присутствующие многомерные векторы в виде упорядоченного по их возрастанию списка и гистограмма. В [4] были определены: мера изолированности для унимодального кластера $m^j(n)$ (1), и мера качества распределения в целом $m(n)$ по $K(n)$ кластерам (2).

Мера (1) определяется как отношение среднего значения гистограммы на границе к максимальному значению гистограммы кластера. Мера (2) измеряет среднюю разделимость кластеров. Минимумы (2) соответствуют лучшим классификациям для различных диапазонов значений n. Всегда $m^j(n)\le 1$ и $m\ (n)\le 1$. Заметим, что границы не являются общими для соседних кластеров, т.к. алгоритм жесткий – кластеры не имеют общих точек. Граница в формуле (1) определяется для каждого кластера, как совокупность точек, таких, что для каждой из них найдется хотя бы одна соседняя точка в списке соседей, не принадлежащая этому же кластеру. Этот список уже найден для алгоритма Нарендры. Для тесно расположенных унимодальных кластеров мера (2) удовлетворяет требованиям, предъявляемым к мерам качества классификации в задачах кластерной достоверности [5]. Природа классифицируемых объектов, – типов покрытия земной поверхности, – такова, что подавляющая часть спектральных признаков составляют кластеры, тесно примыкающие друг к другу. Как показывают исследования, с увеличением числа уровней квантования, наблюдается тенденция к уменьшению средней разделимости кластеров и к росту значения меры (2). Кластеры, не имеющие соседей на границах, большая редкость. Распределение векторов в таких кластерах не гауссово, поэтому компактность кластера не может быть измерена дисперсией, и часто применяемые оценки разделимости, связывающие дисперсию кластеров и расстояние между ними, не подходят. Кроме того, дисперсия становится зависимой от размеров кластеров, и, следовательно, от расстояния между ними. Вообще, гистограммный алгоритм Нарендры имеет смысл применять именно к тесно расположенным кластерам, так как для изолированных кластеров есть более простые методы кластеризации.

Иерархический алгоритм находит сначала число уровней квантования, при котором получается новая система объединенных векторов, такая, что ее унимодальные кластеры наилучшим образом изолированы. Затем внутри каждого полученного кластера алгоритм увеличивает число уровней квантования, и находит свое лучшее кластерное распределение и так далее.

Мера изолированности отдельного кластера (1) не зависит от остальных кластеров, поэтому в качестве меры разделимости распределения в целом по K кластерам для иерархического алгоритма по-прежнему возьмем среднюю разделимость M_K:

$$M_K=\frac{1}{K}\sum_{j=1}^{K}m^j(n^j).\ (3)$$

где n^j теперь число уровней квантования при получении j кластера.

ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

Конечной целью предлагаемого иерархического алгоритма может быть выбор такого общего распределения, при котором разделимость совокупности всех полученных кластеров M_K лучшая. Или такой выбор, при котором разделимость подкластеров внутри каждого кластера не может быть больше заданного порога – точности разделимости ε . При этом в результате оценки разделимости кластеров может быть осуществлен возврат к предыдущему этапу для определенной части данных. По завершении классификации, т.е. достижении n = n1 внутри каждого кластера, в обратном порядке начинается выбор этапа иерархии для каждого кластера, с тем, чтобы уменьшить (3). Последовательно сравниваются значения средней меры разделимости группы кластеров, полученных делением материнского кластера предыдущего этапа иерархии, и меры разделимости самого материнского кластера.

При этом может быть введена фильтрация мелких кластеров. Кластеры, площадь которых меньше вводимого порога (в пикселах), исключаются из суммы (2). Этот шаг оправдан, так как небольшое число элементов этих кластеров не обеспечивает статистической надежности в оценке изолированности кластера (1), гистограмма для них не может быть аппроксимацией плотности вероятности векторов признаков, и применение метод Нарендры не оправдано. Эти кластеры могут быть присоединены к оставшимся по принципу худшей разделимости, или в случае изолированности, по ближайшему расстоянию до центров. Но в обоих случаях оставшиеся кластеры это только те, которые получены делением материнского кластера. Исследования показывают, что таких кластеров может быть много, но общий объем данных в них невелик. Они представляют кластерную пыль. Однако их вклад в сумму (3) может быть существен, если учесть, что мера качества (1) для них обычно велика.

В результате окончательной классификации получается распределение данных по кластерам, представленных с дифференцированной детальностью в соответствии со степенью их изолированности. Наименьший минимум меры обычно достигается для небольшого числа уровней квантования и разделяет векторное пространство на несколько крупных, хорошо разделенных кластеров. С увеличением детализации для каждого внутреннего подкластера, точность повышается.

Окончательная кластеризация гиперспектральных снимков выглядит так, как показано на рис. 2. Здесь 19 кластеров. Областей, нарушающих условие разделимости, нет. Можно получить больше кластеров, если более детально выбрать сетку. Все зависит от конкретной задачи.

Таким образом, видим, что в данном случае (для спутниковых изображений поверхности Земли) алгоритм кластеризации остается таким же, как и для мультиспектральных изображений. Отличие в том, что при сокращении числа спектральных каналов, используется исходное изображение с 220 каналами. Перед использованием алгоритмов гиперспектральное изображеиие приводится в формат гиперспектрального. На рис.2 А разница между кластерами зависит от точности, детальности сетки.

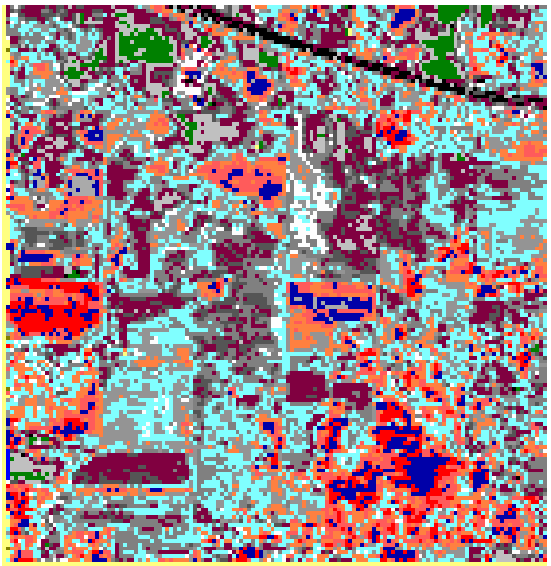

Рис. 2. Кластерная карта гиперспектрального изображения участка поверхности Земли.

Литература:

1. Калиткин Н.Н. Численные методы. Москва. “ Наука ”. 1978. СС. 512.

2. V.S. Sidorova. Detecting Clusters of Specified Separability for Multispectral Data on Various Hierarchical Levels // Pattern Recognition and Image Analysis. - 2014, - Vol. 24, No. 1. – P. 151-155.

3. Narendra P.M. and Goldberg M. A non-parametric clustering scheme for LANDSAT // Pattern Recognition. – 1977 – 9 – P. 207 -215.

4. Сидорова В.С. Оценка качества классификации многоспектральных изображений гистограммным методом // Автометрия. – 2007. – Том 43. – №1. – С. 37- 43.

5. M. Halkidi, Y. Batistakis and M. Vazirgiannis. // Journal of Intelligent Information Systems – 2001 – No.17 (2-3) – P.107-132