UNSUPERVISED CLASSIFICATION OF MULTISPECTRAL IMAGES

Valeria S. Sidorova

Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Novosibirsk, Russia

Abstract

In the proposed hierarchical histogram algorithm, the clustering detail is found, which is different in the subdomains of the vector space of spectral features depending on the given separation of the clusters. The question of reducing the dimension of the space of data attributes is also being considered. The application of the algorithm for uncontrolled classification of terrestrial cover using various satellite remote sensing data is illustrated.

Keywords: remote sensing, uncontrolled classification, multidimensional histogram, cluster separation, dimensionality of space, hierarchical algorithm

НЕКОНТРОЛИРУЕМАЯ КЛАССИФИКАЦИЯ МУЛЬТИСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ

Сидорова В.С.

Институт вычислительной математики и математической геофизики СО РАН, Новосибирск

В предложенном иерархическом гистограммном алгоритме отыскивается детальность кластеризации, различная в подобластях векторного пространства спектральных признаков в зависимости от заданной отделимости кластеров. Также рассматривается вопрос о сокращении размерности пространства признаков данных. Иллюстрируется применение алгоритма для неконтролируемой классификации земного покрова по различным спутниковым данным дистанционного зондирования.

Ключевые слова: дистанционное зондирование, неконтролируемая классификация, многомерная гистограмма, разделимость кластеров, размерность пространства, иерархический алгоритм.

Введение. Предлагается неконтролируемая иерархическая классификация мультиспектральных спутниковых данных с заданием порога отделимости кластеров и сокращением размерности данных.

Внутри кластеров осуществляется кластеризация по унимодальным кластерам методом [1]. Быстрый непараметрический не итеративный алгоритм [1] разделяет векторное пространство признаков по унимодальным кластерам, модальные векторы которых соответствуют локальным гистограммным максимумам, а границы проходят по долинам гистограммы. Он решает три задачи кластеризации одновременно, используя метод графов: находит локальные максимумы гистограммы, объявляя их корнями деревьев-кластеров, проводит границы между кластерами по долинам гистограммы и относит все вектора признаков к своим кластерам. Алгоритм является жестким: каждый вектор принадлежит только одному кластеру. Детальность кластеризации в алгоритме [1] регулируется заданием числа уровней предварительного квантования, одинакового для всего векторного пространства. Задается отсечением младших битов в байтах различных векторных направлений. Каждое такое отсечение соответствует уменьшению векторного пространства вдвое. Алгоритм [1] был реализован автором для ЭВМ "БЭСМ-6", а затем для PC[2,3]. Для получения критерия качества классификации предложено ввести разделимость и отделимость кластеров. В [4] предложен иерархический алгоритм, определяющий детальность в зависимости от кластерной разделимости подобластей. Но иногда бывает важно выявить предельную детальность для данной разделимости. В другом иерархическом алгоритме [5] ставится другая цель: для подобластей пространства векторов найти такие свои максимальные детальности, при которых еще порождаются дочерние кластеры с разделимостью ниже заданной. Это позволяет с одной стороны исследовать структуру многоспектральных данных достаточно подробно, причем на разных иерархических уровнях, с другой стороны, регулировать детальность исследования.

Кроме того, можно по-разному задавать закон изменения детальности. Если различные спектральные направления не эквивалентны, то и менять их можно по-разному. И это может привести к сокращению размерности векторного пространства признаков.

Этапы иерархического алгоритма. Параметром детальности кластеризации является число уровней квантования векторного пространства, обозначим его n. На каждом этапе предложенного иерархического алгоритма находим такое число n (и соответствующие новые векторы g), при котором распределение по унимодальным кластерам, полученное методом [1], дает абсолютный минимум мере (2), в диапазоне изменения $255>n> n1$.

В [6] были предложены: мера отделимости унимодального кластера $m^j(n)$ (1), и мера средней разделимости $K(n)$ кластеров $m(n)$ (2):

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

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

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.

В [6] показано, что (2) удовлетворяет требованиям, предъявляемым к мере разделимости кластеров или качества кластеризации [7,8]. Минимумы (2) соответствуют лучшим классификациям. Всегда $m^j(n) \le 1$ и $m(n) \le 1$. Традиционные меры разделимости оперируют среднеквадратичным отклонением и расстоянием, которые взаимозависимы при жесткой кластеризации. Введенные меры позволят сравнивать изолированность распределений с тесно расположенными кластерами. Более подробно о рассмотренных мерах в [6].

На новом этапе иерархии алгоритм для каждого кластера увеличивает число уровней квантования (детальность), полученное на предыдущем этапе, и в новом интервале находит свое новое число и соответствующее наилучшее кластерное распределение в смысле меры(2). Новый алгоритм рассматривает каждый дочерний кластер как отдельную область только тогда, когда его разделимость удовлетворяет условию (3):

$$m^j(n) < \varepsilon, \qquad (3)$$

где ε заданная точность отделимости кластера.

Плохо разделенные дочерние подкластеры (не удовлетворяющие (3)) рассматриваются как одна область. Детальность квантования увеличивается, и деление продолжается, пока $n<n1$ в каждом дочернем кластере. Затем результаты анализируются снизу-вверх для плохих кластеров, когда условие (3) нарушено. Процесс заканчивается, когда в иерархическом дереве попадется хороший в смысле (3) предок или хорошая сестра. В первом случае данные плохого кластера возвращаются на уровень хорошего предка, во втором кластер считается плохой ветвью и может быть отнесен к общему фону кластеров, не удовлетворяющих (3). Рассмотрим изображение земной поверхности (рис.1); получен со спутника NOAA_17 17 апреля 2003г. Изображение пятиспектральное. Его объем около 1,7 мегабайт, размер 1480*1124 пикселей, разрешение около км* км / пиксель.

Верхнюю левую часть изображения занимают тающие снега тайги Сибири, внизу оттаявшая поверхность Казахстана. Правую часть снимка покрывают сплошные и полупрозрачные облака. Пять спектральных каналов позволяют различать объекты, не различимые лишь в видимом диапазоне. Например, снег и облака визуально не различимы по цвету и яркости, но в диапазоне инфракрасного 3,55 – 3,93 мкм снег сильно поглощает излучение от Земли и выглядит даже черным, а облака белые. В диапазонах 10,3 – 11,3 мкм и 11,5 - 12,5 мкм (рис.1 d и рис.1 e) могут быть различены типы облаков (кучевые, перистые и другие, имеющие разную высоту). Для построения многомерной гистограммы был применен алгоритм с использованием чередования таблиц хеширования и сортировки [9].

Задано $\varepsilon = 0.07$. Было пройдено семь этапов иерархического алгоритма, дальнейшее деление кластеров приводило к нарушению условия разделимости (3) для большой части подкластеров.

Пространственное расположение кластеров на карте хорошо соответствуют земным объектам. Интересно заметить, что три кластера, полученные автоматически новым алгоритмом соответствуют районам добычи полезных ископаемых открытым способом. Они отмечены черным и желтым тонами на карте и локализуются в районах Экибастуза, Баянаула - угледобыча, в Майкаине золото открытым способом; южнее Семипалатинска также уголь и золото. Зоны отечественных разработок, в частности, Кузбасс находятся в заснеженной области под полупрозрачными облаками и не выделились в отдельный кластер. Для всего изображения получено 120 кластеров с отделимостью меньшей $\varepsilon = 0.07$. Для сравнения: при той же макси-

мальной детальности n = 50, одинаковой для всего пятиспектрального изображения алгоритмом [1] получено 2686 кластеров. Время работы нового алгоритма – несколько минут на одноядерном компьютере.

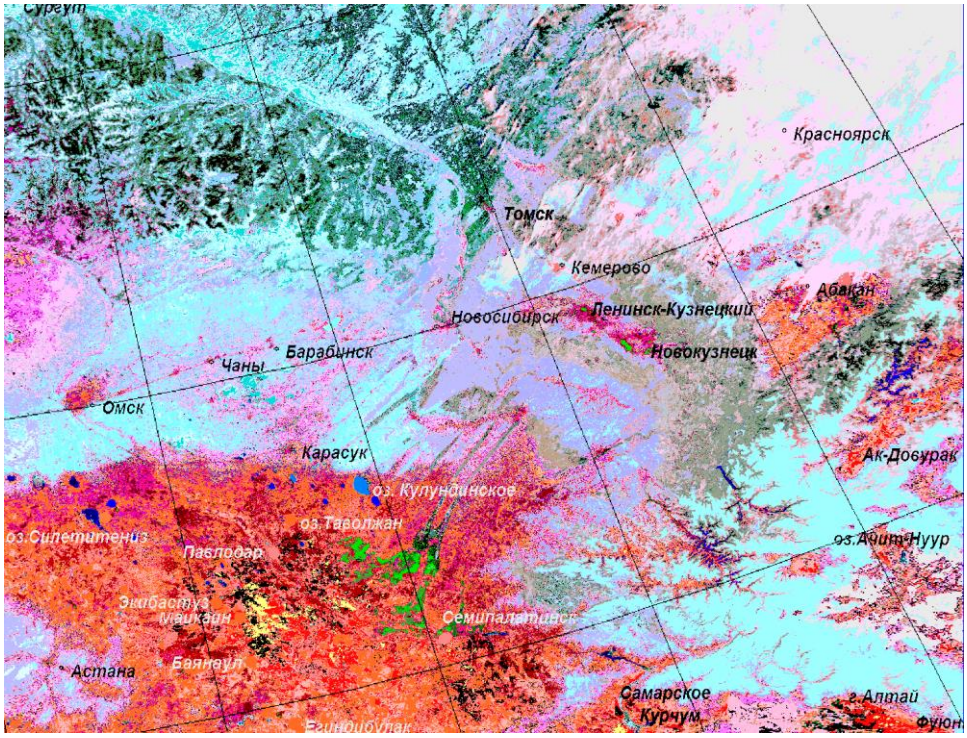

Рис.1. Кластерная карта седьмого этапа иерархии.

Другим важным аспектом является уменьшение размерности данных. Квантование пространства признаков может производиться по разным правилам. До сих пор в каждом спектральном направлении число уровней квантования сохранялось одинаковым. Однако, в общем случае, данные вытянуты вдоль какого-то направления, и 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:

$$\frac{N_{e1}}{S_{e1}} = \frac{N_{e2}}{S_{e2}} = \ldots = \frac{N_{ek}}{S_{ek}} \quad , \tag{1}$$

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

Для перевода пространства в пространство собственных векторов применялся метод Якоби [10]. Зададим максимальное число уровней квантования в собственном пространстве равным N_{em}=255, таково обычное число уровней серого для данных дистанционного зондирования по каждому измерению. 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. Таким образом, если отношение $S_{em}/S_{ex} < 2$, то соответствующая ось x может не рассматриваться, и мы получаем сокращение размерности пространства признаков.

Алгоритм был апробирован на семиканальном (в видимом и инфракрасном диапазонах) изображении Омской области ИСЗ «Landsat-8» 08.02.2014 (размер 3 161×2 590 пикселей, разрешение 15 м), предоставленном сибирским центром ФГБУ «НИЦ «Планета»». Алгоритм кластеризации предварительно осуществляет сокращение размерности векторного пространства спектральных признаков с семи до трех. Детальность, различная по полученным кластерам, определяется делимым иерархическим гистограммным алгоритмом для предельной отделимости кластеров $d = 0{,}15$ ($0 < d < 1$). Иерархичность получающихся кластерных карт отражает иерархичность реальных объектов на космоснимках. Выбор числа этапов осуществляется совместно с экспертом. Например, на первом этапе иерархии получено всего 6 кластеров, два из которых (красный и черный) соответствуют дымам ТЭЦ Омской области. Для десяти этапов иерархии получено 27 унимодальных кластеров (Рис. 2); специалисты НИЦ «Планета» установили, что расположение ярко розового и фиолетового кластеров соответствует загрязнению территории Омской области.

Можно также рассматривать не все пространство векторов при сокращении размерности, а лишь часть его, относящуюся к каждому кластеру. Поскольку алгоритм иерархический, и каждый кластер делится на подкластеры на каждом этапе иерархии в соответствии с изменением детальности представления данных, то размерность каждого подкластера может изменится. При решении задачи внутри кластера (построении ковариационной матрицы) используется уже построенная ранее гистограмма признаков в виде определенным образом организованного списка. Рассмотрим пример

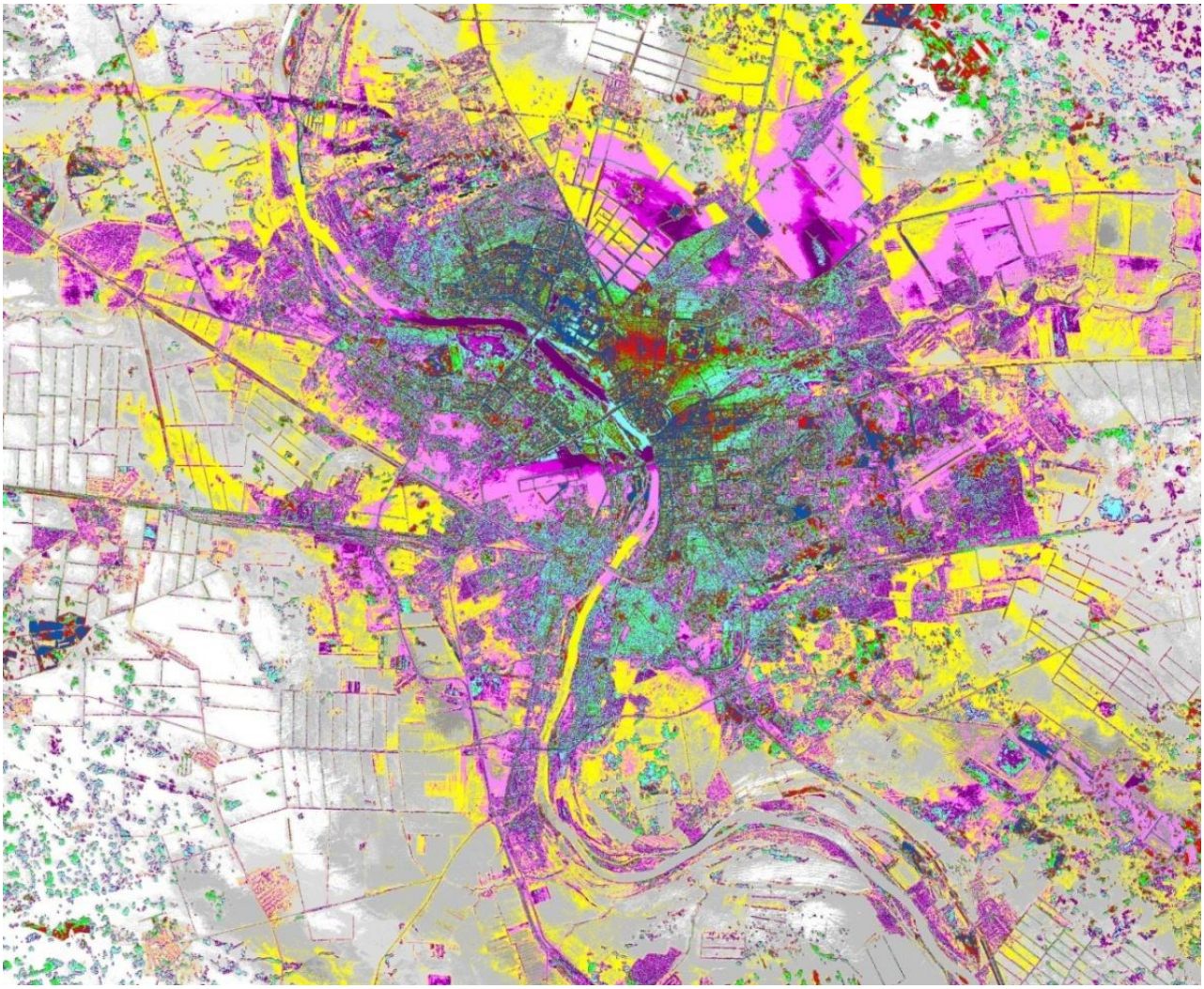

Рис.2. Кластерная карта, полученная делимым иерархическим гистограммным алгоритмом. 15 этапов иерархии. d= 0, 015. Получено 54 кластера (включая маленькие). Загрязнение: лиловые оттенки.

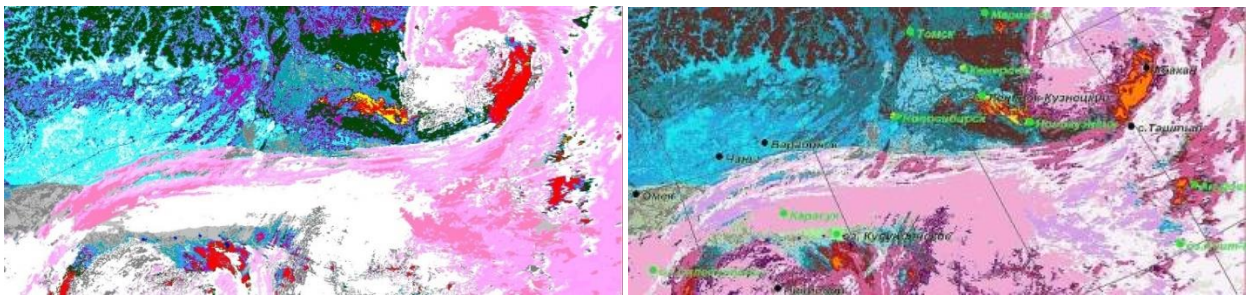

Рис. 3.а) Кластеризация по пяти спектральным каналам иерархическим гистограммным алгоритмом; 5 этапов иерархии; задано *d*=0,1. Получено 22 кластера. б) Кластеризация иерархическим гистограммным алгоритмом с поиском размерности по кластерам; 15 этапов иерархии; задано *d*=0,1; получено 40 кластеров.

Анализируется изображение поверхности Земли со спутника NOAA 17 от 7.04.2003, полный кадр (1328x624) пикселей представлен в пяти спектральных каналах (один в видимой части спектра, остальные в инфракрасной), 4 мегабайт. В нижней части снимка формирование вихря, озера; в верхней в основном – тающие снега, тайга Сибири. У левой границы-Омск, у правой-Абакан. Хорошо прослеживается железная дорога: Омск - Чаны - Барабинск - Новосибирск. На рис.3а кластерная карта при глобальном сокращении размерности. На рис. 3b при покластерном. При большей детальности кластеризации (Рис. 3b.) кластер облаков делятся на унимодальные подкластеры с заданной отделимостью 0,1, и некоторые из них требуют пятиспектрального рассмотрения. Это полупрозрачные облака. Время вычислений оказалось в три раза меньше, чем для полностью пятиспектрального варианта и составило несколько минут на одноядерном компьютере.

Работа выполнена при финансовой поддержке РФФИ (проект № 16-07-00066) и Программы I.33П фундаментальных исследований Президиума РАН (проект № 0315-2015-0012)).

ЛИТЕРАТУРА

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

[2] Сидорова В.С. Кластеризация многоспектральных изображений с помощью анализа многомерной гистограммы // Новосибирск. Сб.: Математические и технические проблемы обработки изображений. СО АН СССР. 1986. С. 52-57.

[3] Сидорова В.С. Классификация многоспектральных космических изображений поверхности Земли с помощью разделения многомерной гистограммы по унимодальным кластерам // Ж. Вестник КазНУ., сер. географическая. 2004. N 2(19). С. 206-210.

[4] V.S. Sidorova. Automatic Hierarchical Clustering Algorithm for Remote Sensing Data. // Pattern Recognition and Image Analysis. 2011. Vol.21. No.2. P. 328 - 331.

[5] 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.

[6] Сидорова В.С. Оценка качества классификации многоспектральных изображений гистограм - мным методом // Автометрия. 2007. Том 43. №1. С. 37- 43.

[7] M. Halkidi, Y. Batistakis and M. Vazirgiannis. On clustering validation techniques // Journal of Intelligent Information Systems. 2001. No.17 (2-3), P.107-132.

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

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

[10] Калиткин Н.Н. Численные методы. Москва. “Наука”. 1978. С. 512