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

Иерархический гистограммный кластерный алгоритм с выбором размерности пространства спектральных признаков для данных дистанционного зондирования Земли[1] Сокращение размерности данных кластерным алгоритмом

Сидорова Валерия Сергеевна, научный сотрудник
Институт вычислительной математики и математической геофизики
Сибирского отделения Российской академии наук (г. Новосибирск)

Если спектральные данные дистанционного зондирования Земли составляют кластеры в какой-либо области спектра, то представляется возможным применить к данным кластерный анализ. Проведенная глобальная сегментация по результатам кластеризации позволит визуально сопоставить полученные области кластеров с известными наземными наблюдениями. Если часть объектов какого-то кластера известна, то кластерный анализ позволит найти и другие его объекты, более точно отследить границы, позволит также “поиграть” с окраской кластеров и более четко их увидеть. Кроме того, кластерный анализ определяет спектральные признаки кластеров как характеристики наиболее плотной (модальной) части кластера, разброс по каждому спектральному каналу, площадь кластеров и др. Кластеризация тем более актуальна, что многоспектральные данные дистанционного зондирования Земли (ДЗЗ) имеют огромный объем.

Кластеризация большого объема данных ДЗЗ обычно осуществляется двумя способами: по К центрам (заранее должно быть известно число кластеров К и приближенное распределение данных) и гистограммными. Здесь предлагается гистограммный алгоритм. Многомерная гистограмма рассматривается как плотность вероятности различных векторов. Трудностью гистограммных алгоритмов является большой объем оперативной памяти, который требуется для хранения гистограммы. Но используемый подход Нарендры [1] сохраняет лишь присутствующие на изображении вектора в виде определенным образом упорядоченного списка. Для ускорения построения гистограммы используется система хэширования. Для хранения векторов и гистограммы было предложено использовать два списка: основной и Shell [2], в котором вектора упорядочены определенным образом (по возрастанию компонент, как в словаре). Идея в том, что, если в основном списке хэш номеров место оказывается занято другим вектором, то этот другой вектор переносится в список Shell, а в основной список вносится новый вектор. Чередование списков при построении гистограммы обеспечивает быстроту поиска и экономит память. После просмотра изображения и построения гистограммы, оба списка объединяются.

Быстрый непараметрический алгоритм Нарендры позволяет полностью автоматизировать процесс кластеризации с целью распознавания данных. Он не требует a priory задания числа кластеров и их формы. Детальность кластеризации задается в нем, однако, произвольно. Она в алгоритме Нарендры определяется предварительным отсечением младших битов в каждом байте, соответствующем спектральному направлению. Хотя идея предварительного квантования пространства признаков, получения новых векторов объединением старых, а затем их кластеризации, обеспечивает получение унимодальных кластеров - и в этом ценность подхода Нарендры.

Отсечение каждого бита соответствует уменьшению числа квантовых уровней вдвое по соответствующему направлению. Такое изменение детальности приводило к резким скачкам в результатах кластеризации. Число полученных кластеров менялось на порядки. Также отметим, что как и большинство кластерных алгоритмов, алгоритм Нарендры не определял разделимость полученных кластеров, хотя качество кластеризации оценивается именно по разделимости кластеров [3]. В первоначальном виде алгоритм Нарендры был реализован автором этой статьи на БЭСМ-6 [4], затем перенесен на персональный компьютер [5].

Учитывая сложность сцены, то есть то, что существует вложенность кластеров: большие кластеры могут быть разделены на подкластеры, а также то, что в одной области данных кластеры наиболее разделены при одной детальности данных, а в другой области при другой, был разработан иерархический делимый гистограммный кластерный алгоритм с заданием отделимости кластеров в приложении к данным ДЗЗ [6]. На каждом этапе иерархии использовался гистограммный алгоритм Нарендры [1], который разделяет пространство признаков по унимодальным кластерам. Новый иерархический алгоритм предлагал автоматизировать процесс выбора максимальной детальности, учитывая разделимость кластеров, причем

[1] Работа выполнена частично при финансовой поддержке Российского фонда фундаментальных исследований (проект № 16-07-00066) и Программы I.33П фундаментальных исследований Президиума РАН (проект № 0315-2015-0012).

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

в различных областях данных разной детальности, обеспечивающей заданную отделимость кластеров *d*.

Для оценки отделимости кластера используется ранее предложенная автором мера [7]. Мера отделимости для отдельного унимодального кластера $m^j(n)$ (1), и мера качества распределения в целом $m(n)$ по $K(n)$ кластерам (2):

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

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

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.

Минимумы меры (2) соответствуют лучшим распределениям с наиболее разделенными кластерами в среднем. Всегда $m^j(n) \le 1$ и $m(n) \le 1$. Ценность этих мер в том, что они позволяют сравнивать распределения с тесно расположенными унимодальными кластерами, когда на их границах много общих векторов. Эти меры удовлетворяют условиям мер отделимости [3]: значения их убывает с увеличением расстояния между кластерами и ростом компактности кластеров (в смысле близости векторов кластера к модальному вектору). Кроме того, эти меры легко вычисляются, так как сравнивают скалярные значения гистограммы в центре и на границах кластеров. Границы кластеров легко находятся, используя списки соседей векторов, построенных как составная часть алгоритма Нарендры.

Устройство иерархического алгоритма следующее. Начиная с малого числа уровней квантования и постепенно его увеличивая, определяется распределение с наилучшим разделением (используя меру (2)) полученных кластеров. Затем на следующем этапе иерархии процесс повторяется внутри каждого полученного кластера, независимо от того, удовлетворяет ли кластер условию, что его отделимость(1) меньше заданного d ($0<d<1$), так как при большей детальности он может разделиться на хорошо разделимые . Начальным значением числа уровней квантования является значение, полученное на предыдущем этапе иерархии для соответствующего кластера плюс один. Однако неудовлетворительные кластеры для дальнейшего деления объединяются в один ложный кластер. Весь иерархический процесс можно прекратить по заданию разных условий: либо числу этапов иерархии, либо из каких-то других физических соображений. Без задания условий он может остановиться, когда в каждом полученном кластере число уровней квантования достигнет заданного (максимально 255). Затем осуществляется автоматический анализ и возврат к тем детальностям, на которых кластеры были разделены по порогу *d*. Таким образом, осуществляется дифференцированный подход к различным областям данных, имеющих различную природу и различную разделимость кластеров. Более подробно алгоритм описан в [6].

Исследованиями установлено, что для данных ДЗЗ с увеличением детальности существует предельная детальность, выше которой разделимость кластеров становится хуже. Таким образом, задавая предельную отделимость кластеров, избегаем получения лишней дробности и получаем существенно меньше кластеров, чем прямым алгоритмом. Причем кластеры хорошо отделены.

Кроме определения детальности по кластерам, можно также дифференцированно (по кластерам) решить вопрос о сокращении числа измерений пространства спектральных признаков по сравнению с исходным (полученным со спутника). Эти два аспекта (число квантовых уровней и размерность собственного пространства) связаны между собой. Анализируем их связь. Квантование пространства признаков может производиться по разным правилам. Ранее 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:

$$\frac{N_{e1}}{S_{e1}} = \frac{N_{e2}}{S_{e2}} = \ldots = \frac{N_{ek}}{S_{ek}}, \qquad (3)$$

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

То есть, правило (3) задает соотношение между числами уровней квантования по координатным осям. Зададим максимальное число уровней квантования в собственном пространстве равным N_{em} =255, таково

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

обычное число уровней серого для данных дистанционного зондирования по каждому измерению. Тогда, в соответствии с пропорциями (3) 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 $\frac{S_{em}}{S_{ex}} < 2$, то соответствующая ось *x* может не рассматриваться, и мы получаем сокращение размерности пространства признаков.

При решении задачи внутри кластера (построении ковариационной матрицы) используется уже построенная ранее гистограмма признаков в виде определенным образом организованного списка. Рассмотрим пример. Анализируется пятиспектральное изображение поверхности Земли со спутника NOAA 17 от 7.04.2003, объем около 4 мегабайт. Спектральные каналы следующие: 1) 0,58-0,68 мкм, 2) 0,725 - 1 мкм, 3) 3, 55 - 3,93 мкм, 4) 10,3-11,3 мкм, 11) 11,5-12,5 мкм. На рис.1 представлено изображение в одном из каналов с показом некоторых пунктов для ориентировки. В нижней части снимка формирование вихря, озера; в верхней в основном – тающие снега, тайга Сибири.

Кластеры с площадью меньше 100 пикселей отнесены в фоновый. На рис.3 карта для семи этапов иерархии, *d*=0,12. Получено 29 кластеров. Все кластеры унимодальные в своей области детальности. Фонового кластера нет. Данные кластеров, соответствующих суши, имеют размерность, равную трем, облака в основном размерность, равную четырем, но полупрозрачные облака требуют пятиспектрального рассмотрения.

Число кластеров оказалось меньше, чем для варианта без сокращения размерности. Время вычислений оказалось в три раза меньше, чем для пятиспекрального варианта и составило несколько минут на одноядерном компьютере РК 1.6 ГГц 512 МБ.

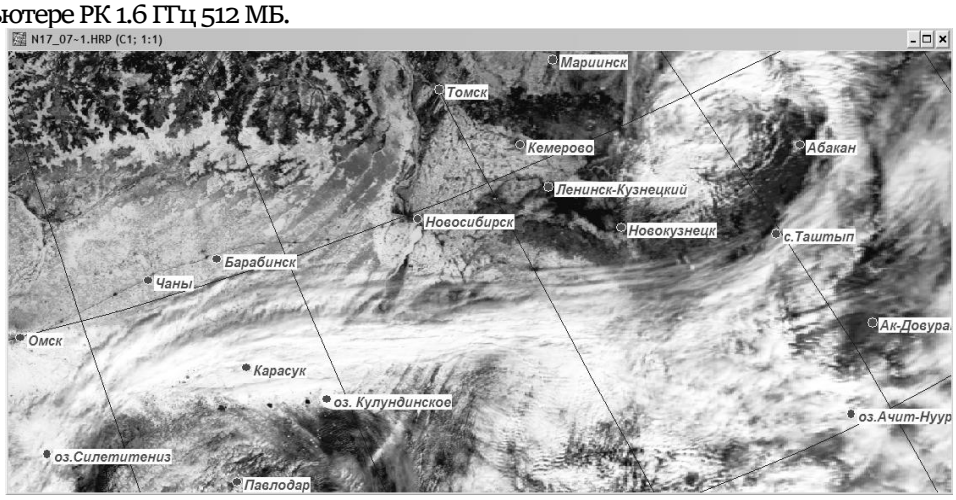

Рис. 1. **Исходное изображение в видимой части спектра.**

Рис. 2. **Кластеризация иерархическим гистограммным алгоритмом с поиском размерности по кластерам;**
7 этапов иерархии; задано *d*=0,12; получено 29 кластеров

Другой пример. Здесь размерность была сокращена перед использованием алгоритма кластеризации. На рис. 3 представлено изображение района Улан-Удэ; цель - выделение области загрязнения территории. Это семиспектральное изображение Бурятии со спутника “Landsat-8. Исходный файл предоставлен сибир-

www. esa-conference.ru

ским центром ФГПУ НИЦ “ПЛАНЕТА”. Построение ковариационной матрицы спектральных данных для всего изображения и ее диагонализация методом Якоби [8] показало, что можно рассматривать три измерения без существенной потери информации. Сокращение размерности приводит к экономии компьютерного времени. К преобразованным данным был применен делимый иерархический гистограммный алгоритм кластеризации с порогом отделимости кластера d = 0, 015. Затем проведена глобальная сегментация, и полученная карта унимодальных кластеров представлена на рис.4.

Кластеры с фиолетовыми и темно-зелеными оттенками соответствуют загрязненным территориям, рассчитаны их площади, координаты, модальные вектора и др. Таким образом, исходный семиспектральный файл объемом около 36 Мбайт был разделен на 54 унимодальных кластера, каждый с отделимостью не хуже d= 0, 015 за несколько минут на одноядерном компьютере.

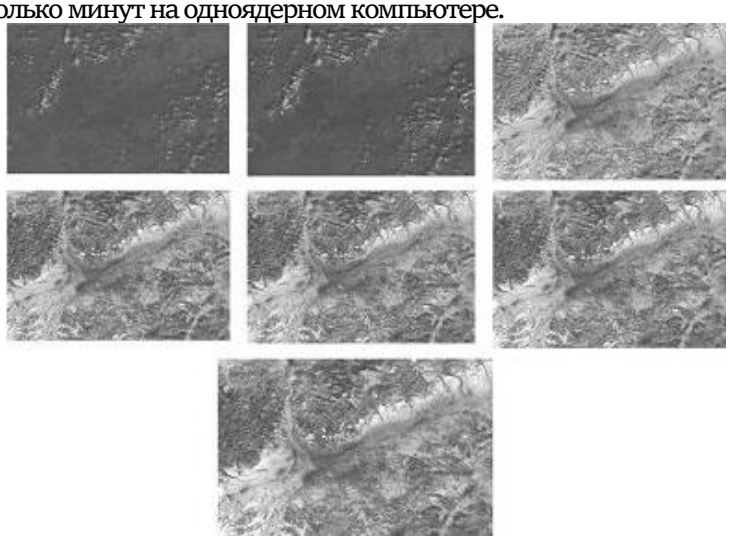

Рис. 3. **Семиспектральное изображение со спутника “Landsat-8”, район Улан-Удэ**

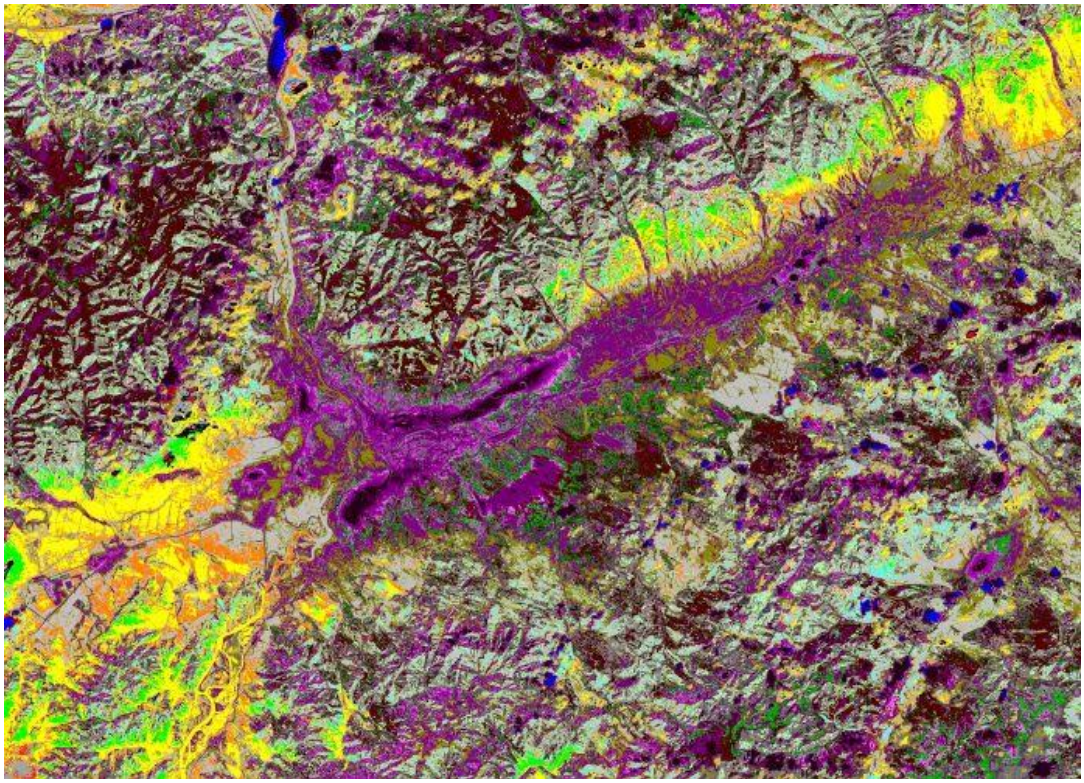

Рис.4. **Кластерная карта, полученная делимым иерархическим гистограммным алгоритмом. 15 этапов иерархии. d= 0, 015. Получено 54 кластера (включая маленькие вплоть до 1 пикселя). Загрязнение: лиловые и темно-зеленые оттенки**

Литература:

1.Narendra P.M. and Goldberg M. A non-parametric clustering scheme for LANDSAT // Pattern Recognition. –

www. esa-conference.ru

1977 – 9 – P. 207 -215.

2. V. S. Sidorova. Separating of the Multivariate Histogram on the Unimodal Clusters. // Proceedings of the Second IASTED International Conference “Automation Control and Information Technology”. – Novosibirsk. – 2005. – P. 267–274.

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

4. Сидорова В.С. Кластеризация многоспектральных изображений с помощью анализа многомерной гистограммы // Новосибирск. Сб. Математические и технические проблемы обработки изображений. СО АН СССР. – 1986 –. СС. 52-57.

5. 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) –. СС. 206-210.

6. 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.

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

8. Калиткин Н.Н. Численные методы. Москва. “ Наука ”. 1978. С. 512.