Journal of Siberian Federal University. Engineering & Technologies, 2016, 9(7), 972-978
~ ~ ~

УДК 528.8.04, 528.88

# Hard and Fuzzy Clustering of the Earth Remote Sensing Data

**Vasiliy V. Asmus[a],**
**Aleksey A. Buchnev[b] and Valeriy P. Pyatkin*[b]**
*[a]State Research Center of Space Hydrometeorology «Planeta», Rosgidromet*
*7 Bolshoi Predtechenskii Str., Moscow, 123242, Russia*
*[b]Institute of Computational Mathematics*
*6 Akademika Lavrenteva, Novosibirsk, 630090, Russia*

Received 09.02.2016, received in revised form 27.04.2016, accepted 19.07.2016

*The clustering system for processing of the Earth remote sensing data is discussed. The system consists of the next methods: K-means method, method of the multidimensional histograms modes analysis, hybrid method, which involves method of the multidimensional histograms modes analysis and the subsequent hierarchical grouping, and a number of fuzzy clustering algorithms.*

*Keywords: remote sensing, clustering, hard clustering, fuzzy clustering.*

Citation: Asmus V.V., Buchnev A.A., Pyatkin V.P. Hard and fuzzy clustering of the earth remote sensing data, J. Sib. Fed. Univ. Eng. technol., 2016, 9(7), 972-978. DOI: 10.17516/1999-494X-2016-9-7-972-978.


* Corresponding author E-mail address: pvp@ooi.sscc.ru

# Жесткая и нечеткая кластеризация данных дистанционного зондирования Земли

**В.В. Асмус[а], А.А. Бучнев[б], В.П. Пяткин[б]**
*[а]Научно-исследовательский центр*
*космической гидрометеорологии «Планета»*
*Россия, 123242, Москва, Большой Предтеченский переулок, 7*
*[б]Институт вычислительной математики*
*и математической геофизики СО РАН*
*Россия, 630090, Новосибирск, пр. Академика Лаврентьева, 6*

*Рассматривается система кластеризации данных дистанционного зондирования Земли (ДЗЗ). Система представлена следующими методами: методом К-средних, методом анализа мод многомерных гистограмм, гибридным методом, объединяющим метод анализа мод многомерных гистограмм с последующей иерархической группировкой и рядом алгоритмов нечеткой кластеризации.*

*Ключевые слова: дистанционное зондирование, кластерный анализ, жесткая кластеризация, нечеткая кластеризация.*

## Введение

Центральные вопросы тематической обработки (интерпретации) данных ДЗЗ – вопросы повышения качества дешифрирования – непосредственно связаны с проблемой выбора адекватных алгоритмов распознавания [1-4]. В данной статье рассматривается кластерный анализ в обработке многоспектральных (многомерных) данных ДЗЗ. Характеризуя методы кластеризации в целом, следует отметить, что в основном они отыскивают в данных не те структуры, которые там реально существуют, а те, для поиска которых они предназначены [2]. Поэтому надежность результатов кластеризации часто можно оценить лишь сравнением нескольких вариантов обработки данных ДЗЗ. Характерная особенность данных ДЗЗ – “загрязнение” выборок смешанными векторами измерений, т.е. векторами, которые образуются при попадании в элемент разрешения съемочной системы нескольких природных объектов. Это обстоятельство является одним из источников ошибок при построении карты классификации [1, 2]. Большинство алгоритмов кластеризации для отнесения векторов признаков кластерам вычисляют для каждого вектора значения подходящей функции «правдоподобия». В случае зачисления вектора признаков в кластер по максимальному значению функции правдоподобия получается так называемая *жесткая* кластеризация. Рассмотрим некоторые алгоритмы *жесткой* кластеризации.

## Жесткая кластеризация

В состав программного комплекса, реализованного совместными усилиями ФГБУ «НИЦ «Планета» и ФГБУН ИВМиМГ СО РАН, входит реализация классического алгоритма жесткой кластеризации – алгоритма *K*-средних, широко используемого для разбиения на кластеры больших объемов многомерных данных [4]. Алгоритм *K*-средних может быть отнесен к клас-

су параметрических, так как он неявным образом предполагает природу плотности вероятности: кластеры стремятся иметь конкретную геометрическую форму, зависящую от выбранной метрики. Мы используем следующие метрики: Евклидова, Махаланобиса, Чебышева, city-block-расстояние. Известно также, что результат кластеризации методом *K*-средних зависит от задания начальных центров кластеров. Предоставляется выбор одного из трех вариантов, два из которых определяются на основе статистических характеристик набора данных и один основан на случайной выборке. Один из вариантов алгоритма позволяет учитывать влияние смешанных векторов [2]. Дополнительный параметр в этом случае – выбираемое эмпирически соотношение чистых и смешанных векторов в наборе данных. На основе этого соотношения и градиентного изображения, сформированного подходящим градиентным оператором (Робертса/Превитта/Собела), выделяются связные компоненты, состоящие из чистых векторов. Кластеризации подвергаются средние векторы связных компонент. В дальнейшем смешанные векторы распределяются по полученным кластерам на основе минимального расстояния до центра кластера.

Другой подход, позволяющий получать разбиение векторов измерений на кластеры произвольной формы, основан на предположении, что исходные данные являются выборкой из многомодового закона распределения, причем векторы, отвечающие отдельной моде, образуют кластер [2]. Таким образом, задача сводится к анализу мод многомерных гистограмм.

Еще один алгоритм жесткой кластеризации, реализованный в нашем программном комплекс, – гибридный метод: анализ мод многомерной гистограммы с последующей иерархической группировкой. Практическое использование метода анализа мод многомерной гистограммы показывает, что зачастую получение приемлемого результата – весьма трудоемкий процесс и требует высокой квалификации эксперта-исследователя. Причиной этого служит, вероятно, то, что алгоритм многопараметрический (в частности, на решение оказывает большое влияние способ сглаживания гистограммы). В связи с этим система кластеризации дополнена двухэтапной процедурой (с сохранением всех ранее существовавших функций): на первом этапе выполняется предварительное разбиение исходной выборки на кластеры с помощью модального анализа, а затем для получения окончательного результата используется иерархическая группировка [5]. Заметим, что применение иерархической группировки для кластеризации исходного набора векторов нереально из-за того, что используемая в алгоритме матрица расстояний состоит (в начале работы алгоритма) из $N(N-1)/2$ элементов, где $N$ – количество векторов. Предварительное использование модального анализа позволяет сократить объем данных до разумных пределов. В качестве входных данных для иерархической группировки используются векторы средних группы векторов, связанных с каждой модой многомерной гистограммы. Напомним, что на каждом шаге восходящей иерархической классификации объединяются два кластера, расстояние между которыми минимально. Достоинством иерархической группировки является то, что после построения иерархического дерева кластеризации можно “разрезать” его на любом уровне иерархии, т.е. получать разные кластерные карты, не запуская снова процесс кластеризации.

Последний этап работы всех алгоритмов жесткой кластеризации – сортировка полученных кластеров по убыванию их объемов и подсчет соответствующих статистик: объемов, векторов средних и девиаций (стандартных отклонений) в каналах для каждого кластера. Эти данные

записываются при необходимости в файл на диске. Туда же записывается число векторов данных, не вошедших ни в один из кластеров, т.е. попавших в "$K$+1"-й кластер. Эти данные служат основой для анализа разделимости полученных кластеров. Результат работы классификаторов в рабочем режиме – одноканальное (байтовое) изображение, значениями пикселов которого являются номера кластеров. Это изображение окрашивается в предопределенные цвета, которые в интерактивном режиме могут быть заменены на цвета, определяемые пользователем. К выходному изображению можно применить функцию постклассификации для удаления изолированных пикселов (генерализация данных).

Приведенные ниже рисунки демонстрируют работу жесткой кластеризации алгоритмом $K$-средних. На рис. 1 приведен фрагмент изображения бассейна Обского водохранилища, полученного 19.04.2011 ИСЗ Terra (EOS AM-1), сканер Modis. Рисунок 2 содержит изображения кластеров (всего их 5), соответствующих состоянию водно-ледовой поверхности водохранилища.

Следует отметить, что предложенные алгоритмы жесткой кластеризации внедрены в практику оперативной работы ФГБУ «НИЦ «Планета» и широко используются в технологии построения тематических карт состояния природных объектов по спутниковым данным видимого, инфракрасного или микроволнового диапазонов.

### Нечеткая (*мягкая)* кластеризация

Альтернативой жесткой разделяющей кластеризации выступает *мягкая,* или *нечеткая*, кластеризация, разрешающая векторам принадлежать всем кластерам с коэффициентом членства $u_{ij} \in [0,1]$, определяющим степень принадлежности $j$-го вектора $i$-му кластеру:

$$\sum_{i=1}^{C} u_{ij} = 1, \ \forall j, \qquad (1)$$

$$\sum_{j=1}^{L} u_{ij} < L, \ \forall i,$$

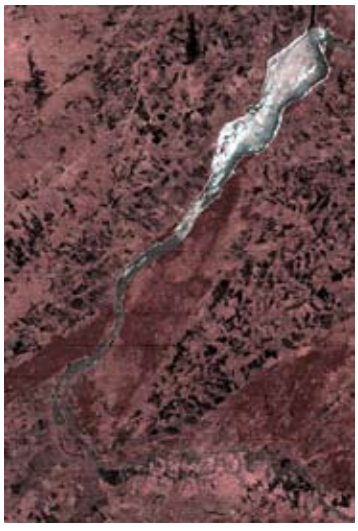

Рис. 1

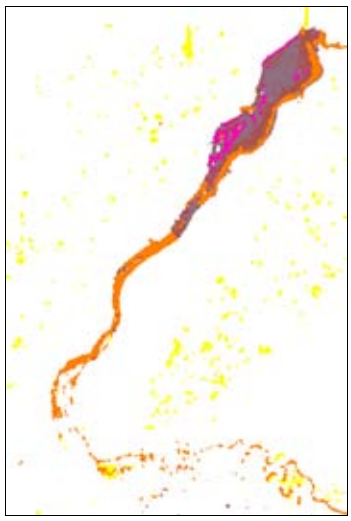

Рис. 2

определяя этими соотношениями нечеткую кластеризацию. Здесь $C$ – число кластеров, $L$ – количество векторов измерений. В недавнее время нами в состав системы кластеризации программного комплекса была включена реализация широко используемого алгоритма нечеткой кластеризации, известного как метод *C-средних* [6]. Это итерационный алгоритм, который используется для разделения смешанных векторов измерений в данных ДЗЗ. Идея метода заключается в описании сходства вектора с каждым кластером с помощью функции уровней принадлежности, принимающей значения от нуля до единицы. Значения функции, близкие к единице, означают высокую степень сходства вектора с кластером. Очевидно, что сумма значений функции уровней принадлежности для каждого пиксела должна равняться единице. Как и в алгоритме *K*-средних, параметрами соответствующей процедуры (кроме количества кластеров) служат тип метрики и вариант выбора начальных центров кластеров. Дополнительным параметром является показатель нечеткости, значения которого для данных ДЗЗ предлагается брать близкими к двум (см. [1]).

На рис. 3 представлено изображение, полученное ИСЗ «Метеор-М1» 03.04.2012. Рисунок 4 демонстрирует результат работы алгоритма *C*-средних (выделялось 20 кластеров).

Вторым алгоритмом нечеткой кластеризации, включенным в состав программного комплекса по обработке данных ДЗЗ, является алгоритм нечеткой кластеризации с регуляризацией – так называемый алгоритм *Possibilistic C-means, PCM*. Принципиальное отличие алгоритма PCM от алгоритма FCM состоит в снятии ограничения (1) на элементы матрицы принадлежности вектора признаков кластерам: в алгоритме FCM для каждого вектора признаков сумма элементов матрицы принадлежности по всем кластерам должна равняться единице (вероятностное – probabilistic – свойство алгоритма FCM). Таким образом, в алгоритме FCM членство вектора в кластере относительно, так как оно зависит от членства этого вектора во всех других кластерах, в то время как в алгоритме PCM значение членства вектора в кластере абсолютно (т.е. не зависит от значений членства этого вектора в других кластерах) и может интерпретироваться в терминах типичности вектора. Алгоритм PCM пытается найти моды в наборе данных, так как каждый полученный кластер соответствует плотной области в этом наборе. В процес-

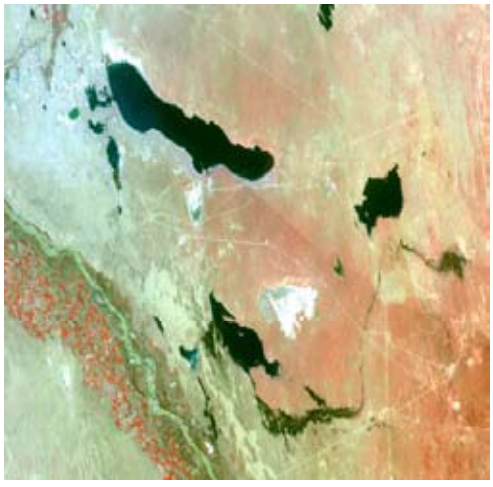

Рис. 3

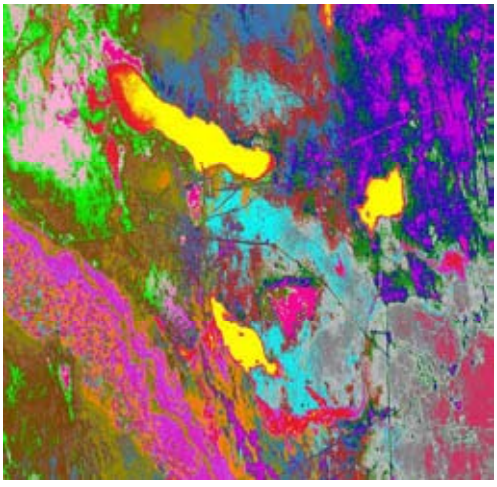

Рис. 4

се выполнения итераций алгоритма прототипы кластеров последовательно перемещаются в плотные области в пространстве признаков.

PCM-алгоритм является робастным методом кластеризации, который может быть использован для обнаружения плотных областей в данных. Степень членства вектора признаков в кластере определяется двумя величинами: расстоянием вектора до прототипа кластера и параметром $K$, называемым ссылочным расстоянием кластера. Значение этого параметра индивидуально для каждого кластера и зависит от среднего размера кластера.

Авторы алгоритма (Krishnapuram&Keller [7]) отмечают, что для получения качественных результатов кластеризации требуется хорошая инициализация ссылочных расстояний кластеров. Следуя их рекомендациям, в качестве начального приближения матрицы степеней членства векторов признаков в кластерах используют результат выполнения алгоритма нечеткой кластеризации методом FCM. Другими словами, необходимым условием выполнения алгоритма PCM для какого-либо набора данных является предварительное выполнение алгоритма FCM для этого набора данных.

Основная часть работы алгоритмов FCM и PCM состоит в итерационном перестроении матрицы уровней принадлежности векторов признаков кластерам и пересчете центров кластеров. Алгоритмы заканчивают работу при выполнении заданного числа итераций либо при достижении матрицы уровней принадлежности состояния стабильности, т.е. состояния, при котором норма разности матриц в двух последовательных итерациях не превосходит заданного порога. Эта работа требует больших временных затрат при ее последовательном выполнении, особенно в случае, когда показатель нечеткости не равен двум, в связи с чем реализованы параллельные версии алгоритмов. Параллельная реализация алгоритмов осуществляется средствами ОС Windows в рамках одного процесса путем запуска нескольких параллельных потоков. Количество запускаемых потоков равно количеству логических процессоров компьютера. Каждый поток перестраивает соответствующую часть матрицы уровней принадлежности. Необходимая при работе параллельных потоков синхронизация достигается с помощью механизма событий ОС Windows.

## Заключение

Практика решения конкретных прикладных задач ДЗЗ с использованием предлагаемых алгоритмов кластеризации многоспектральных космических снимков, получаемых как с российских, так и с зарубежных спутников, подтверждает их высокую эффективность. Отметим, что широкий набор возможностей системы кластеризации программного комплекса позволяет эксперту-исследователю выбирать адекватные решения задач дешифрирования данных ДЗЗ.

***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).***

**Список литературы**

[1] Шовенгердт Р.А. *Дистанционное зондирование. Модели и методы обработки изображений.* М.: Техносфера, 2010. 560 с. [Schowengerdt R.A. Remote sensing. Models and methods for image processing. Moskva, Technosfera, 2010, 560 p. (in Russian)]

[2] Асмус В.В. *Программно-аппаратный комплекс обработки спутниковых данных и его применение для задач гидрометеорологии и мониторинга природной среды.* Докторская диссертация. Москва, 2002. 75 с. [Asmus V.V. Software-hardware complex processing of satellite data and its application to problems of hydrometeorology and environmental monitoring. Doctoral thesis, Moskva, 2002. 75 p. (in Russian)]

[3] Асмус В.В., Бучнев А.А., Пяткин В.П. Кластерный анализ данных дистанционного зондирования Земли. *Автометрия*, 2010, 46(2) , 58-66. [Asmus V.V., Buchnev A.A., Pyatkin V.P. The cluster analysis of Earth remote sensing data, *Avtometriya*, 2010, 46(2), 58-66. (in Russian)]

[4] Jain A.K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010, 31. 651-666.

[5] М. Жамбю. *Иерархический кластер-анализ и соответствия*. М.: Финансы и статистика, 1988. 342 с. [M. Jambu. Hierarchical cluster analysis and compliance. Moskva, Finansy I statistika, 1988. 342 p. (in Russian)]

[6] Bezdek J.C. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981. 163 p.

[7] R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering . *IEEE Transactions on Fuzzy Systems*, 1993, 1, 98–110.