

УДК 528.8.04, 528.88

Fuzzy Clusters with Volume Prototypes in the Thematic Processing of the Earth Remote Sensing Data

Aleksey A. Buchnev and Valeriy P. Pyatkin*

*Institute of Computational Mathematics
and Mathematical Geophysics SB RAS
6 Akademika Lavrentieva, Novosibirsk, 630090, Russia*

Received 26.01.2017, received in revised form 14.03.2017, accepted 19.07.2017

The fuzzy clustering technology of the Earth remote sensing data, based on extended C-means and Gustafson-Kessel algorithms, is discussed. The algorithms extensions consist of clusters with volume prototypes construction and using of clusters similarity measure. The volume prototypes are less sensitive to a bias in the distribution of the data, and similar clusters are merged during clustering.

Keywords: fuzzy clustering, volume prototypes, clusters similarity, extended C-means algorithm, extended Gustafson-Kessel algorithm.

Citation: Buchnev A.A., Pyatkin V.P. Fuzzy clusters with volume prototypes in the thematic processing of the earth remote sensing data, J. Sib. Fed. Univ. Eng. technol., 2017, 10(6), 723-726. DOI: 10.17516/1999-494X-2017-10-6-723-726.

© Siberian Federal University. All rights reserved

* Corresponding author E-mail address: baa@ooi.sbcc.ru, pvp@ooi.sbcc.ru

Нечеткие кластеры с объемными прототипами в тематической обработке данных дистанционного зондирования Земли

А.А. Бучнев, В.П. Пяткин

*Институт вычислительной математики и
математической геофизики СО РАН*

Россия, 630090, Новосибирск, пр. Академика Лаврентьева, 6

Рассматривается технология нечеткой кластеризации данных дистанционного зондирования Земли (ДЗЗ) расширенными алгоритмами C -средних и Густафсона-Кесселя. Расширения алгоритмов состоят в использовании объемных прототипов и меры сходства кластеров. Объемные прототипы менее чувствительны к шумовым выбросам в распределении данных. Кроме того, использование меры сходства позволяет объединять кластеры в процессе кластеризации.

Ключевые слова: нечеткая кластеризация, объемные прототипы, сходство кластеров, расширенный алгоритм C -средних, расширенный алгоритм Густафсона-Кесселя.

Одним из основных вопросов тематической обработки (интерпретации) данных ДЗЗ является вопрос повышения качества распознавания. Возникающие трудности обусловлены, в частности, «загрязнением» выборок смешанными векторами признаков, т.е. векторами, которые образуются при попадании в элемент разрешения съемочной системы нескольких природных объектов [1, 2]. Большинство алгоритмов классификации для отнесения векторов признаков классам вычисляют для каждого вектора значения подходящей функции «правдоподобия». В случае зачисления вектора признаков в класс по максимальному значению функции правдоподобия получается так называемая жесткая кластеризация (например, так работает классический алгоритм K -средних).

Альтернативным подходом служит нечеткая кластеризация, которая позволяет каждому вектору признаков принадлежать одновременно всем кластерам с определенной степенью членства (принадлежности) в каждом кластере. Одним из первых алгоритмов нечеткой кластеризации является, вероятно, алгоритм C -средних [3]. Вопросы реализации и использования этого алгоритма в обработке данных ДЗЗ освещаются в [4].

Дальнейшим развитием системы нечеткой кластеризации данных ДЗЗ выступает реализация нечеткой кластеризации расширенными алгоритмами C -средних (*Fuzzy C-means – FCM*) и Густафсона-Кесселя (*Gustafson-Kessel – GK*) [5]. В алгоритме *FCM* выбранная метрика, определяющая форму получаемых кластеров, одинакова для всех кластеров и не меняется в процессе работы. Принципиальное отличие алгоритма *GK* от алгоритма *FCM* состоит в том, что каждый кластер имеет индивидуальную метрику, основанную на нечеткой ковариационной матрице кластера (метрика Махаланобиса). Эта метрика динамически меняется в процессе выполнения итераций алгоритма.

Расширения *FCM*- и *GK*-алгоритмов (получаются *E-FCM*- и *E-GK*-алгоритмы) состоят в следующем [5].

В качестве прототипов кластеров используются объемные прототипы (*volume prototypes*). Объемный прототип – выпуклое и компактное подпространство пространства признаков. В частности, если в алгоритме *E-FCM* используется евклидова метрика, тогда таким прототипом будет гипершар с центром v_i и радиусом r_i для i -го кластера. Вектор признаков x_k , для которого расстояние $d_{ik}(x_k, v_i) \leq r_i$, является элементом объемного прототипа и входит в i -й кластер с членством 1.0. В алгоритме *E-GK* объемный прототип кластера – гиперэллипсоид.

Мера расстояния, используемая в расширенных алгоритмах, определяется на основе исходной меры d . Для вектора x_k сначала вычисляется значение $d_{ik}(x_k, v_i)$, а затем расстояние до объемного прототипа берется равным $\widetilde{d}_{ik} = \max(0, d_{ik} - r_i)$. Таким образом, векторы признаков, лежащие внутри объемного прототипа i -го кластера, имеют нулевое значение членства в других кластерах, что уменьшает влияние плотных областей в данных на центры других кластеров. Размеры объемных прототипов определяются на основе объемов кластеров.

Вводится понятие «сходства» (*similarity*) кластеров. Сходство кластеров определяется на основе нечеткой включающей меры. Для двух нечетких кластеров $u_i(x_k)$ и $u_j(x_k)$ нечеткая включающая мера определяется как

$$I_{ij} = \frac{\sum_{k=1}^N \min(u_{ik}, u_{jk})}{\sum_{k=1}^N u_{ik}}.$$

Здесь N – число векторов признаков в наборе данных; u_{ik} – значение членства k -го вектора в i -м кластере. Включающая мера не является симметрической, но она используется для построения симметрической меры сходства нечетких кластеров $u_i(x_k)$ и $u_j(x_k)$:

$$S_{ij} = \max(I_{ij}, I_{ji}).$$

Начиная с заведомо большего числа кластеров, кластеры, степень сходства которых превышает заданный порог, объединяются в итерационном процессе кластеризации для того, чтобы получить подходящее разбиение данных.

Приведенные ниже рисунки демонстрируют процесс построения карты нечеткой кластеризации алгоритмом *E-FCM*. На рис. 1 приведен фрагмент изображения бассейна Обского водохранилища, полученного 19.04.2011 ИСЗ Terra (EOS AM-1), сканер Modis. Рис. 2 содержит изображения нечетких кластеров (всего их два), соответствующих состоянию водно-ледовой поверхности бассейна водохранилища.

Заметим, что во входных данных было задано 7 кластеров, в результате работы алгоритма все векторы признаков были разбиты по двум кластерам. В качестве начального разбиения векторов признаков по нечетким кластерам используются выходные данные алгоритма *C-средних*.

Соответствующая вычислительная процедура реализована в рамках модели параллельного программирования OpenMP под управлением ОС Windows. Включение алгоритмов *E-FCM* и *E-GK* в состав программного комплекса по обработке спутниковых данных PlanetaMonitoring позволяет расширить возможности системы кластеризации комплекса по построению разбиения, наиболее полно соответствующего внутренней структуре данных.



Рис. 1
Fig. 1

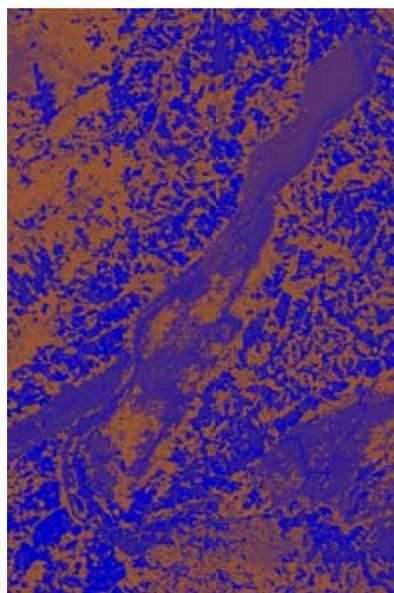


Рис. 2
Fig. 2

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] Асмус В.В. *Программно-аппаратный комплекс обработки спутниковых данных и его применение для задач гидрометеорологии и мониторинга природной среды*. Москва, 2002. 75 с. [Asmus V.V. Software-hardware complex processing of satellite data and its application to problems of hydrometeorology and environmental monitoring, Moskva, 2002. 75 p. (in Russian)]
- [2] 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)]
- [3] Bezdek J.C. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981.
- [4] Асмус В.В., Бучнев А.А., Пяткин В.П. Жесткая и нечеткая кластеризация данных дистанционного зондирования Земли. *Журнал СВУ. Сер. Техника и технологии*, 2016, 9(7), 972-978. DOI: 10.17516/1999-494X-2016-9-7-972-978. [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. (in Russian)]
- [5] Uzay Kaimak and Magne Setnes. *Extended Fuzzy Clustering Algorithms*. ERIM report series ERS-2000-51-LIS. Rotterdam, Netherlands, November 2000, 24.