FUZZY CLASSIFICATION OF THE EARTH REMOTE SENSING DATA

Alexey A. Buchnev, Valeriy P. Pyatkin

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

Abstract

The system of fuzzy classification of Earth remote sensing (ERS) data is discussed. The system involves fuzzy automatic classification (clustering) and fuzzy supervised classification. The fuzzy clustering subsystem consists of the next algorithms of fuzzy clustering: fuzzy C-means (FCM), fuzzy C-means with regularization (PCM) and extended C-means and Gustafson-Kessel algorithms. The fuzzy supervised classification subsystem involves the Wang's method and the explicit fuzzy supervised classification method.

Keywords: remote sensing, clustering, hard clustering, fuzzy clustering, probabilistic fuzzy clustering, possibilistic fuzzy clustering, supervised classification, fuzzy supervised classification

НЕЧЕТКАЯ КЛАССИФИКАЦИЯ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ

Бучнев А.А., Пяткин В.П.
Институт вычислительной математики и математической геофизики СО РАН

Рассматривается система нечеткой классификации данных дистанционного зондирования Земли (ДЗЗ). Система включает нечеткую автоматическую классификацию (кластеризацию) и нечеткую контролируемую классификацию. Подсистема нечеткой кластеризации включает реализацию следующих алгоритмов: алгоритма С-средних (FCM), алгоритма С-средних с регуляризацией (PCM) и расширенных алгоритмов С-средних и Густафсона-Кесселя. Подсистема нечеткой контролируемой классификации включает метод Вонга и метод явной нечеткой контролируемой классификации.

Ключевые слова: дистанционное зондирование, кластерный анализ, жесткая кластеризация, нечеткая кластеризация, вероятностная нечеткая кластеризация, возможностная нечеткая кластеризация, контролируемая классификация, нечеткая контролируемая классификация.

Введение. Характерной особенностью данных ДЗЗ является “загрязнение” выборок смешанными векторами признаков, 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]. Большинство алгоритмов классификации для отнесения векторов признаков кластерам (классам) вычисляют для каждого вектора значения подходящей функции «правдоподобия». В случае зачисления вектора признаков в кластер (класс) по максимальному значению функции правдоподобия получается так называемая *жесткая* классификация.

Альтернативой жесткой разделяющей классификации является *мягкая* или *нечеткая* классификация, разрешающая векторам измерений принадлежать всем кластерам (классам) с коэффициентом членства $u_{ij} \in [0,1]$, определяющим степень принадлежности *j*-го вектора *i*-му кластеру (классу):

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

$$0 < \sum_{j=1}^{L} u_{ij} < L, \ \forall i,$$

определяя этими соотношениями нечеткую классификацию. Здесь C – число кластеров (классов), L – количество векторов признаков.

Нечеткая кластеризация. В недавнее время нами в состав подсистемы кластеризации программного комплекса по обработке данных ДЗЗ была включена реализация широко используемого алгоритма нечеткой кластеризации, известного как метод *С-средних* (*Fuzzy C-means, FCM*) [2]. Это итерационный алгоритм, который используется для разделения смешанных векторов признаков в данных ДДЗ. 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. Здесь сумма значений функции уровней принадлежности для каждого пиксела равняется единице. Параметрами соответствующей процедуры (кроме числа кластеров) являются тип метрики и вариант выбора начальных центров кластеров. Дополнительным параметром является показатель нечеткости, значения которого для ДДЗ предлагается брать близкими к двум (см., например, [1]).

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 является относительным, т.к. оно зависит от членства этого вектора во всех других кластерах, в то время как в алгоритме PCM значение членства вектора в кластере является абсолютным (т.е. не зависящим 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 процессе 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.

Нижеследующие рисунки демонстрируют результаты работы алгоритмов C-средних. На рис. 1 представлен фрагмент снимка ИСЗ SPOT-4, полученного 04.05.2011 г., с паводковой ситуацией в районе Камня-на-Оби (снимок предоставлен Сибирским центром НИЦ «Планета»). На рис. 2 приведен результат обработки алгоритмом FCM. Фрагменты исходного изображения, являющиеся «шумом» по отношению к области интереса, исключены из процесса обработки. На рис. 3 и 4 представлены результаты обработки алгоритмом PCM со значениями ссылочных расстояний $K=1$ и $K=0.8$ соответственно. Выделялось 10 кластеров, выполнялось 50 итераций алгоритмов.

Авторы алгоритма [3] 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 используется результат выполнения алгоритма нечеткой кластеризации методом FCM. Т.е. необходимым условием выполнения алгоритма PCM для какого-либо набора данных является предварительное выполнение алгоритма FCM для этого набора данных.

Рис. 1. Исходное изображение.

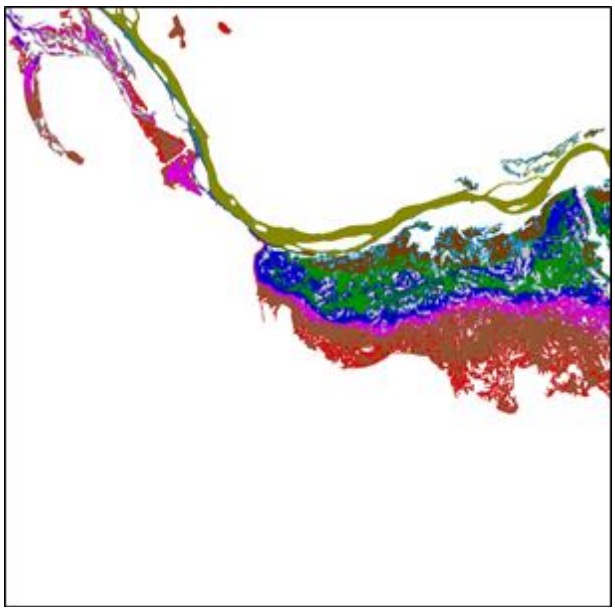

Рис. 2. Кластеризация методом FCM.

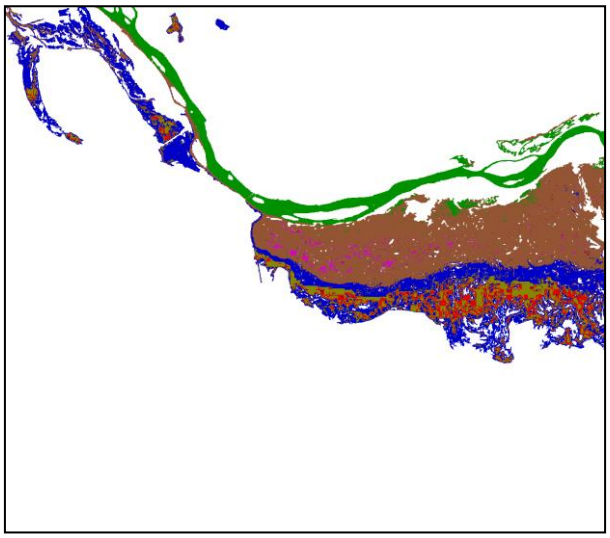

Рис. 3. Кластеризация методом PCM с *K=1.*

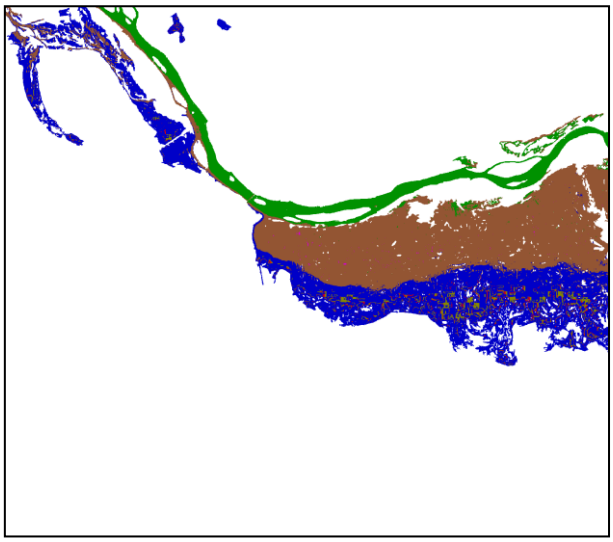

Рис. 4. Кластеризация методом PCM с *K=0.8.*

Дальнейшим развитием системы нечеткой кластеризации данных ДЗЗ является реализация нечеткой кластеризации расширенными алгоритмами С-средних (*Fuzzy C-means – FCM*) и Густафсона-Кесселя (*Gustafson-Kessel – GK*) [4]. 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:

1. В качестве прототипов кластеров используются объемные прототипы (*volume prototypes*). В частности, если в алгоритме *E-FCM* используется евклидова метрика, тогда таким прототипом будет гипершар. В алгоритме *E-GK* объемным прототипом кластера является гиперэллипсоид. Размеры объемных прототипов определяются на основе объемов кластеров. Такие прототипы менее чувствительны к отклонениям в распределении данных.
2. Вводится понятие «сходства» (*similarity*) кластеров. 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.

Заметим, что в качестве начального разбиения векторов признаков по нечетким кластерам используются выходные данные алгоритма *C*-средних.

Основная часть работы алгоритмов нечеткой кластеризации 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 временн’ых затрат при ее последовательном выполнении, особенно в случае, когда показатель нечеткости неравен 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 реализуется с

помощью механизма событий ОС Windows. В таблице 1 содержатся данные о времени выполнения параллельной процедуры нечеткой кластеризации методом FCM набора векторов признаков рис. 1. Приводятся результаты измерений времени (в секундах) для значений параметра нечеткости m=2 и m=2.2. Измерения проводились под управлением Windows-10 на аппаратной платформе i3-2100 с четырьмя логическими процессорами. Выполнялось 50 итераций. Аналогичные данные для алгоритма PCM приведены в таблице 2.

Нечеткая контролируемая классификация. Вонг [1] изменил традиционный метод максимального правдоподобия путем предварительного вычисления нечеткой ковариационной матрицы. Затем степени нечеткого членства векторов в классах вычисляются путем применения процедуры максимального правдоподобия к нечетким сигнатурам классов.

В общем случае алгоритм Вонга требует априорных знаний о членствах в классах векторов из обучающих выборок. Мы в своей реализации метода используем для этого выходные данные алгоритма нечеткой кластеризации [2].

Таблица 1. Время работы алгоритма FCM.

Значение m	Количество запускаемых потоков			
	1	2	3	4
m=2	76.18	52.87	44.52	40.65
m=2.2	305.56	189.14	140.04	116.92

Таблица 2. Время работы алгоритма PCM.

Значение m	Количество запускаемых потоков			
	1	2	3	4
m=2	50.71	33.03	30.95	28.80
m=2.2	228.60	123.51	99.12	82.66

С другой стороны, существует метод нечеткой контролируемой классификации, не требующий никаких априорных сведений о нечеткой принадлежности классам векторов из обучающих выборок. Соответствующий алгоритм известен как алгоритм явной нечеткой контролируемой классификации [5]. Этот алгоритм включает три основных шага. Сначала выполняется преобразование векторов признаков в нечеткое представление (*fuzzification step*) для получения оценок вкладов классов в каждую спектральную полосу при условии гауссова распределения векторов признаков в классах. Фактически это преобразование сводится к вычислению для компонент каждого вектора признаков значений нормального распределения в соответствующей спектральной полосе во всех классах:

$$\boldsymbol{f_{b,c}(x_b)} = \exp\left(-\frac{\left(\boldsymbol{x_b} - \boldsymbol{\mu_{b,c}}\right)^2}{2\boldsymbol{\sigma}^{*2}_{\boldsymbol{b,c}}}\right).$$

Здесь $\boldsymbol{x_b}$ – компонента вектора признаков в спектральной полосе b, $b = 1,\ldots,B$, а параметры нормального распределения определяются с помощью сигнатур классов (сигнатуры классов строятся на основе обучающих выборок): среднее значение $\boldsymbol{\mu_{b,c}}$ совпадает с оценкой среднего для класса c, $c=1,\ldots,C$, в спектральной полосе b, а стандартное отклонение $\boldsymbol{\sigma^*_{b,c}}$ является модулированным значением оценки стандартного отклонения $\boldsymbol{\sigma_{b,c}}$. Значение коэффициента модуляции определяется на основе ожидаемого размера класса в данной полосе. Таким образом, на первом этапе с каждым вектором признаков связывается матрица F, число столбцов которой равно числу классов, а число строк равно размерности вектора признаков:

$$\boldsymbol{F} = \begin{pmatrix} \boldsymbol{f_{1,1}(x_1)} & \boldsymbol{f_{1,2}(x_1)} & \cdots & \boldsymbol{f_{1,C}(x_1)} \\ \boldsymbol{f_{2,1}(x_2)} & \boldsymbol{f_{2,2}(x_2)} & \cdots & \boldsymbol{f_{2,C}(x_2)} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{f_{B,1}(x_B)} & \boldsymbol{f_{B,2}(x_B)} & \cdots & \boldsymbol{f_{B,C}(x_B)} \end{pmatrix}.$$

Эти матрицы являются входными данными ко второму этапу. На втором этапе к полученным данным применяется правило нечеткого вывода для получения, после нормирования,

нечеткой классификации набора векторов признаков. В качестве правила нечеткого вывода используется одно из двух правил Мамдани: правило *MIN* для получения минимального значения в списке аргументов и правило *PRODUCT* для получения произведения значений аргументов. Эти правила применяются к столбцам матриц *F*. Наконец третий этап алгоритма, используя правило *MAX* для выбора среди смешанных в векторе признаков тематических классов класса с максимальным значением членства, переводит нечеткую классификацию в жесткую (*defuzzification step*).

Заключение. Включение алгоритмов нечеткой классификации в состав системы тематической обработки программного комплекса по обработке данных ДЗЗ позволяет построить карту классификации, более полно соответствующую истинным тематическим классам в наборе данных.

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

ЛИТЕРАТУРА

[1] Шовенгердт Р.А. Дистанционное зондирование. Модели и методы обработки изображений. Пер. с англ. Москва: Техносфера, 2010.

[2] Bezdek J.C. Pattern recognition with fuzzy objective function algorithms. N.Y.: Plenum Press, 1981.

[3] Krishnapuram R., Keller J.M. A possibilistic approach to clustering // IEEE Trans. on Fuzzy Systems. 1993. Vol. 1. P. 98–110.

[4] Kaimak U., Setnes M. Extended Fuzzy Clustering Algorithms: ERIM report series ERS-2000-51-LIS. Rotterdam, Netherlands, November 2000. 24 p.

[5] Melgani F., Al Hashemy B., Taha S. An Explicit Fuzzy Supervised Classification Method for Multispectral Remote Sensing Images // IEEE Trans. on Geosci. and Remote Sens. Jan. 2000. Vol. 38, N 1. P. 287-295.