

ІНФАРМАТЫКА

УДК 004

Б. А. ЗАЛЕССКИЙ, Э. Н. СЕРЕДИН

МЕТОД НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ k -СРЕДНИХ СО ШТРАФНОЙ ФУНКЦИЕЙ
И ПРОИЗВОЛЬНЫМИ СТЕПЕНЯМИ ВЕКТОРА ПРИНАДЛЕЖНОСТИ

Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь,
e-mail: zalesky@newman.bas-net.by, eduard.seredin@tut.by

Представлена новая версия метода нечеткой кластеризации k -средних, названная методом нечеткой кластеризации k -средних со штрафной функцией. Разработан алгоритм кластеризации полутоновых, цветных и мультиканальных изображений. Особенностью алгоритма является использование аддитивной штрафной функции гиббсовского типа для контроля локальной гладкости на однородных областях кластерного представления изображений. Локальная гладкость получаемых решений зависит от значений коэффициентов штрафной функции. Эти коэффициенты могут быть взяты как функции величины и направления градиента изображения. В отличие от известного алгоритма нечеткой кластеризации k -средних, который неустойчив по отношению к шумам и искажениям, присутствующим на изображении, предлагаемый алгоритм позволяет строить устойчивые кластерные представления на зашумленных и поврежденных изображениях. Свойства алгоритма проиллюстрированы на тестах с реальными изображениями. Эксперименты показали, что алгоритм, с одной стороны, корректно объединяет однородные области изображений в отдельные кластеры, а с другой – предотвращает слияние разнородных областей.

Ключевые слова: нечеткая кластеризация, k -средних, штрафная функция, вектор принадлежности, множители Лагранжа, EM-алгоритм, GPFCM, CUDA.

B. A. ZALESKY, E. N. SEREDIN

FUZZY C-MEANS CLUSTERING METHOD WITH THE GIBBS PENALTY FUNCTION
AND ARBITRARY POWERS OF THE VECTOR ACCESSORIES

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus,
e-mail: zalesky@newman.bas-net.by, eduard.seredin@tut.by

A new version of the Fuzzy C-means clustering method, which we have called the Fuzzy C-means with the Gibbs Penalty (GPFCM), is introduced. The algorithm is designed to cluster gray scale, color and multichannel images. The specificity of the algorithm is to use the additive Gibbs penalty function in order to control the local smoothness of cluster representations of images in the regions of image homogeneity. The local smoothness of the solutions found by the GPFCM depends on the values of the coefficients of the penalty function. In turn, these coefficients can be taken as functions of magnitude and direction of the image gradient. Unlike the standard Fuzzy C-means clustering algorithm, which is unstable with respect to image noise and distortions, the proposed GPFCM allows building robust cluster approximations of noisy and corrupted pictures. Algorithm properties are illustrated on real images. Experiments have shown that on the one hand, the algorithm properly represents the homogeneous image regions as separate clusters and, on the other hand, it prevents the merging of different homogeneous areas.

Keywords: fuzzy clustering, C-means, penalty function, vector accessories, Lagrange multipliers, EM algorithm, GPFCM, CUDA.

Введение. В настоящее время различные версии методов кластеризации k -средних являются популярным инструментом для обработки многомерных числовых данных и изображений различной природы и модальности. Особенно это относится к методу нечеткой кластеризации k -средних, который избавлен от некоторых недостатков его предшественника – классического метода k -средних.

Метод кластеризации k -средних (КМ), предложенный в 1950-х гг. Г. Штейнгаузом [1] и С. Ллойдом [2], был развит рядом авторов, среди которых Дж. Данн, Э. Распини, Дж. Бездек и др.

Последний, основываясь на введенном понятии, построил новый метод, который в дальнейшем стал называться методом нечеткой кластеризации k -средних (FCM). Подробная историческая справка по данному вопросу приведена, например, в работе [3].

При использовании КМ изначально фиксируется число кластеров k и предполагается, что каждый вектор данных (или пиксел изображения) может принадлежать ровно одному кластеру [1, 4]. Действие алгоритма состоит в минимизации среднеквадратичного отклонения (или отклонения в другой метрике) векторов кластера от его центра масс. Первоначально в качестве центров выбираются произвольные k векторов данных, например случайным образом. Каждое наблюдение относится к ближайшему центру, затем центры получившихся кластеров пересчитываются. Векторы данных разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе к вектору в выбранной метрике. Итерации повторяются до момента выполнения выбранного критерия завершения работы алгоритма.

Недостатком КМ является его сходимость (в общем случае) к локальному экстремуму. На практике это приводит, например, к тому, что при кластеризации изображений метод может объединить несколько неоднородных областей в один кластер, даже если их число меньше разрешенного максимального значения k . Помимо этого, из-за сходимости к локальному экстремуму целевой функции нередко результаты кластеризации методом КМ оказываются зависимыми от выбора первого набора центров кластеров.

Одним из основных отличий FCM от классического КМ является то, что первый метод, основанный на понятиях нечеткой кластеризации, ставит в соответствие каждому вектору данных не одно число – номер кластера, а вектор принадлежности, который характеризует «степень» принадлежности вектора данных каждому классу. Такой подход позволяет независимо от начальных значений центров кластеров достигать глобального экстремума целевой функции. Часто вектор принадлежности каждого элемента выборки понимается как набор вероятностей, задающих шансы элемента выборки принадлежать каждому из кластеров.

Известно большое количество публикаций, посвященных различным модификациям FCM [5–7]. Многообразие версий метода объясняется тем, что они приспособлены к кластеризации данных, обладающих определенными свойствами, т. е. не существует версии FCM, которая работала бы одинаково эффективно для всех типов данных.

Кластеризация изображений имеет свою специфику. Дело в том, что изображения характеризуются не только значениями яркости или цвета, но и их расположением в пространстве. Иными словами, с помощью одного и того же набора цветов можно «нарисовать» совершенно разные сцены. Естественно, что в таком случае имеет смысл учитывать пространственные характеристики изображений.

Ранее одним из авторов [8] был разработан метод нечеткой кластеризации k -средних с аддитивной штрафной функцией гиббсовского типа GPFCM, который позволил улучшить качество и надежность кластеризации изображений, особенно при наличии на них шумов и искажений. Однако в указанном методе GPFCM степень, в которую возводились компоненты функции принадлежности пикселов изображений кластерам, была равна 2. Оказалось, что при других значениях степени разработанный алгоритм вычисления оптимального решения в общем случае к нему не сходится.

В данной статье рассмотрен метод GPFCM с произвольными степенями $\gamma > 1$ функций принадлежности. Построен алгоритм, который гарантированно сходится к оптимальному решению. Приведены результаты сравнения GPFCM с методом FCM Бездека. Описана зависимость свойств кластерных представлений изображений от значений степени γ , в которую возводятся координаты функций принадлежности. Показано, что GPFCM, в отличие от КМ и FCM, позволяет получить удовлетворительные кластерные представления зашумленных изображений, более того, что указанный метод может быть использован не только для кластеризации, но и для фильтрации зашумленных изображений.

Процесс построения кластерных представлений с помощью GPFCM требует большего объема вычислений по сравнению с КМ и FCM, однако разработанные параллельные программные реализации метода с помощью технологии программирования CUDA дают возможность получать решения за несколько секунд.

Свойства алгоритма проиллюстрированы с помощью реальных изображений и космических снимков.

1. Описание метода GPFCM со степенями компонент вектора принадлежности $\gamma > 1$. Так как наши результаты связаны с кластеризацией изображений, введем обозначения их элементов. Пусть I – d -канальное изображение (причем число каналов может быть равным 1) размера $m \times n$, которое индексируется пикселями $p = (x, y)$. Каждый элемент изображения I_p представляет собой вектор $I_p = (I_{1,p}, \dots, I_{d,p})^T$ (здесь T – символ транспонирования), координаты которого $I_{i,p}$ равны яркостям i -го канала ($1 \leq i \leq d$).

Предположим, что заранее заданное число кластеров равно k . Центры кластеров обозначим через $c_j = (c_{1,j}, \dots, c_{d,j})^T$, ($1 \leq j \leq k$), а множество всех центров – через C . Буква U будет использоваться для множества $U = \{u_p\}$ векторов принадлежности $u_p = (u_{p,1}, \dots, u_{p,k})$ пикселей p кластерам. Векторы принадлежности должны удовлетворять двум следующим условиям:

$$u_{p,j} \geq 0, \sum_{j=1}^k u_{p,j} = 1 \quad (1)$$

для всех пикселей p . В принятых обозначениях целевая функция метода FCM Бездека может быть записана в виде

$$F(U, C) = \sum_{j=1}^k \sum_p u_{p,j}^\gamma \|I_p - c_j\|^2,$$

где значение степени γ , $\gamma > 1$, выбрано заранее, а $\|\cdot\|$ обозначает выбранную норму, которая в данной работе будет считаться евклидовой.

Ранее в работе [8] функция $F(U, C)$ была модифицирована с целью получения возможности контроля локальной гладкости кластерных представлений изображений. К ней была добавлена штрафная функция гиббсовского типа так, что усовершенствованная целевая функция приобрела вид

$$Q(U, C) = F(U, C) + \sum_{j=1}^k \sum_p \left(\mu_{p,x} (u_{p,j} - u_{p+(1,0),j})^2 + \mu_{p,y} (u_{p,j} - u_{p+(0,1),j})^2 \right), \quad (2)$$

где $\mu_{p,x}$, $\mu_{p,y}$ – неотрицательные коэффициенты, отвечающие за локальную гладкость кластерных представлений, пиксел $p + (0,1)$ – нижний сосед пиксела p , а $p + (1,0)$ – его правый сосед. Коэффициенты $\mu_{p,x}$, $\mu_{p,y}$ могут выбираться в зависимости от градиента $G_p = (G_{p,x}, G_{p,y})$ изображения, например в виде

$$\mu_{p,x} = \begin{cases} \mu, & \text{если } |G_{p,x}| \leq \tau \\ \frac{\mu}{|G_{p,x}| + 1}, & \text{иначе} \end{cases} \quad \text{и} \quad \mu_{p,y} = \begin{cases} \mu, & \text{если } |G_{p,y}| \leq \tau \\ \frac{\mu}{|G_{p,y}| + 1}, & \text{иначе} \end{cases}, \quad (3)$$

где константа $\mu \geq 0$ задает степень гладкости кластерного представления, а порог τ уменьшает зависимость от градиента, образованного шумом. В некоторых случаях применения GPFCM выбор постоянных значений коэффициентов $\mu_{p,x} = \mu_{p,y} = \text{const}$ может дать лучший результат по сравнению с использованием (3). Для задания градиента можно брать операторы Собеля, Превитта или иные [4].

Метод FCM строит кластерное представление на основе условной минимизации целевой функции $F(U, C)$ при условии (1). Традиционно условная минимизация $F(U, C)$ выполняется с помощью градиентного алгоритма пошаговой минимизации [7], который дает хорошие результаты за приемлемое время. Задача условной минимизации $Q(U, C)$ несколько сложнее. Присутствие штрафной функции в целевой функции $Q(U, C)$ приводит к невозможности использования упомянутой схемы, поэтому некоторые шаги алгоритма приходится реализовывать иным способом.

2. Алгоритм нахождения условного минимума целевой функции. Напомним, что кластерное представление в методе GPFCM строится путем условной минимизации целевой функции (2) при условии (3).

С помощью метода множителей Лагранжа задача условной минимизации сводится к безусловной минимизации функции

$$\begin{aligned} \Phi_\gamma(U, C, \Lambda) = & \sum_{j=1}^k \sum_p u_{p,j}^\gamma \|I_p - c_j\|^2 + \\ & + \sum_{j=1}^k \sum_p \left(\mu_{p,x} (u_{p,j} - u_{p+(1,0),j})^2 + \mu_{p,y} (u_{p,j} - u_{p+(0,1),j})^2 \right) + \sum_p \lambda_p \left(1 - \sum_{j=1}^k u_{p,j} \right), \end{aligned}$$

в которой векторная переменная Λ образована множителями Лагранжа λ_p . При $\gamma = 2$ условие $0 \leq u_{p,j} \leq 1$ выполняется автоматически. При других $\gamma > 1$ предлагается использовать измененный градиентный алгоритм, в котором решение вычисляется с соблюдением условия неотрицательности функций принадлежности $u_{p,j}$, так как традиционная схема дает неудовлетворительные результаты. Ниже приведено его пошаговое описание, позволяющее вычислять решения для $\gamma = 2$, а после него представлены изменения вычислений, которые нужно сделать, чтобы алгоритм работал при всех $\gamma > 1$.

Шаг 1. Присвоить начальные значения функциям принадлежности $U = \{u_{p,j}\}$ для всех пикселей p и всех кластеров $1 \leq j \leq k$ с помощью датчика псевдослучайных чисел, равномерно распределенных на отрезке $[0,1]$ так, чтобы было выполнено условие $\sum_{j=1}^k u_{p,j} = 1$.

Шаг 2. Вычислить центры кластеров по формуле

$$c_j = \frac{\sum_p u_{p,j}^\gamma I_p}{\sum_p u_{p,j}^\gamma}, \quad j = 1, \dots, k.$$

Шаг 3. Вычислить значения множителей Лагранжа для всех p :

$$\lambda_p = \frac{\gamma}{k} \sum_{j=1}^k u_{p,j}^{\gamma-1} \|I_p - c_j\|^2,$$

которые получаются дифференцированием целевой функции Φ_γ по переменным $u_{p,j}$ для нахождения минимума $\Phi_\gamma(U, C, \Lambda)$ по Λ при фиксированных значениях U и C с учетом условия

$$\sum_{j=1}^k u_{p,j} = 1.$$

Шаг 4. Обновить значения целевой функции $u_{p,j}$ по формуле

$$u_{p,j} = \frac{\frac{1}{2} \lambda_p + \mu_{p,x} u_{p+(1,0),j} + \mu_{p,y} u_{p+(0,1),j} + \mu_{p-(1,0),x} u_{p-(1,0),j} + \mu_{p-(0,1),y} u_{p-(0,1),j}}{\|I_p - c_j\|^2 + \mu_{p,x} + \mu_{p,y} + \mu_{p-(1,0),x} + \mu_{p-(0,1),y}}. \quad (4)$$

Шаг 5. Повторять шаги 2–4 до момента первого выполнения условия останова.

В качестве условий останова алгоритма используются: а) выполнение заданного числа итераций; б) малые изменения значений функций U принадлежности, оцененные в какой либо норме, например ℓ_q , где $1 \leq q \leq \infty$ или иной норме; в) отсутствие изменений в нескольких полученных последовательных кластерных представлениях или другие.

Предлагаемое пошаговое описание разработанного алгоритма для вычисления кластерного представления изображений (условной минимизации Φ_γ) при $\gamma = 2$ не позволяет получать решения для $\gamma \neq 2$, однако после внесения изменений в шаг 4 алгоритм становится применим для любых $\gamma > 1$. Поэтому в шаге 4 вместо точной формулы (4) нужно использовать неградиентный метод минимизации Φ_γ по $u_{p,j}$, например дихотомии или золотого сечения, для нахождения минимума Φ_γ по $u_{p,j}$ на интервале $[0,1]$.

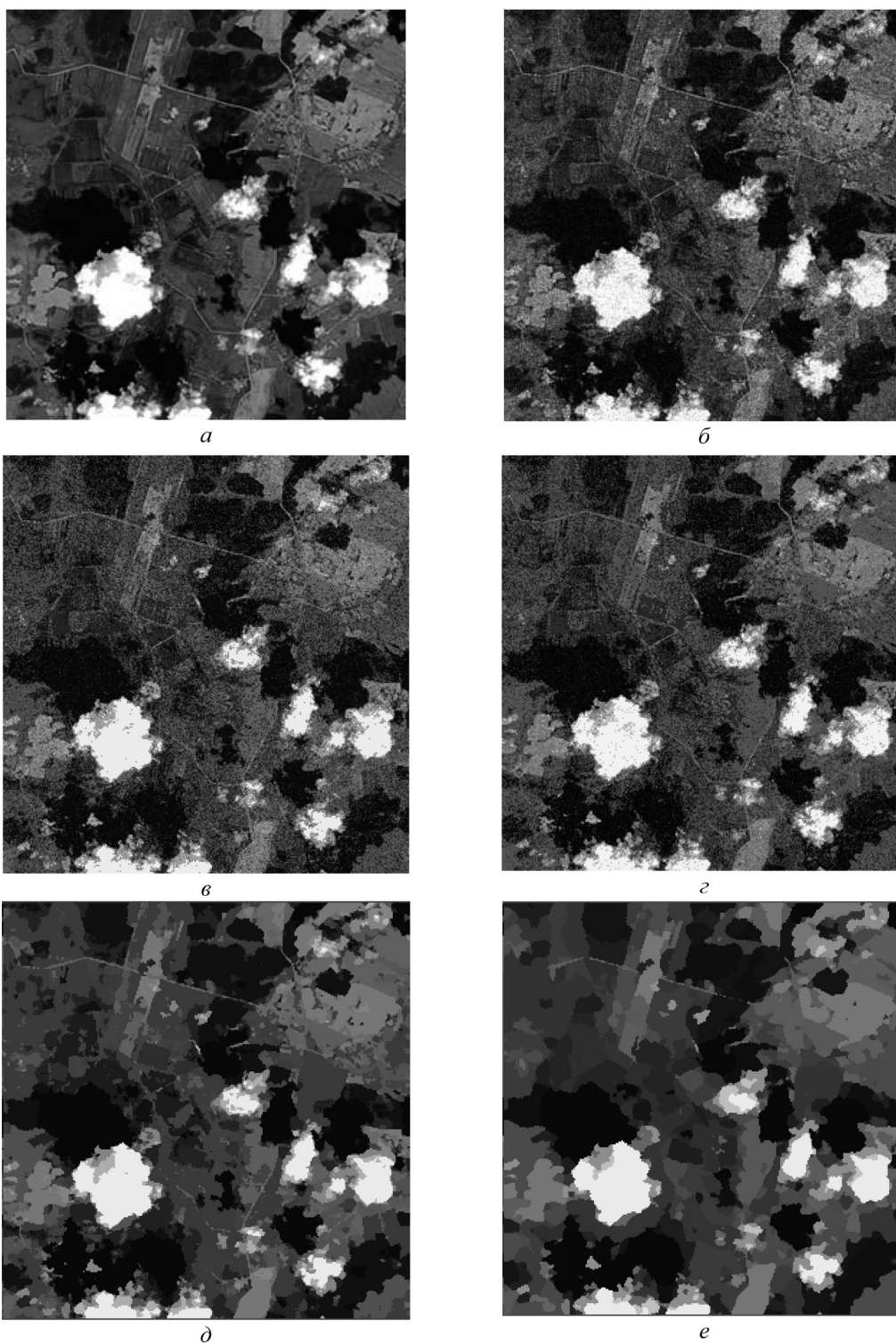


Рис. 1. Оригинальное спутниковое изображение, полученное с помощью алгоритма повышения разрешения (паншарпенинга) (а); оригинальное изображение, искаженное гауссовским случайным шумом (б); результаты кластеризации зашумленного изображения: методом КМ (в); методом FCM (г); методом GPFCM с параметром $\mu = 1000$ (д); методом GPFCM с параметром $\mu = 5000$ (число классов = 20, гамма = 1,5) (е)

Непосредственное вычисление измененной версии алгоритма занимает в несколько раз больше времени, чем вычисление решения методом FCM Бездека, так как вместо точной формулы (4) применяется итерационный метод минимизации φ_γ по $u_{p,j}$. Непосредственное вычисление кластерных представлений больших изображений на CPU может занимать минуты или даже часы,



Рис. 2. Оригинальное изображение с пейзажем (а); оригинальное изображение, искаженное гауссовским случайным шумом (б); результаты кластеризации зашумленного изображения: методом КМ (в); методом FCM (г); методом GPFCM с параметром $\mu = 1000$ (д); методом GPFCM с параметром $\mu = 5000$ (число классов = 20, гамма = 1,5) (е)

однако GPFCM тоже может быть распараллелен как КМ и FCM. Разработанные параллельные программные версии метода на CUDA выполняются всего за несколько секунд. Ниже представлены результаты численных экспериментов.

3. Результаты экспериментов. Для исследования свойств предложенного метода GPFCM были выполнены несколько серий экспериментов на изображениях со спутника Landsat 8 и на фотографиях с пейзажем, искаженных шумами разных типов.

Эксперименты с выбранными изображениями позволили исследовать свойства кластерных представлений, получаемых с помощью GPFCM, и сравнить их со свойствами кластерных представлений, получаемых методами КМ и FCM. Результаты экспериментов подтвердили неустойчивость КМ и FCM к шумам и искажениям, которые разбивали однородные области зашумленных изображений на множество мелких кластеров. Помимо этого проявились известные недостатки КМ, нередко объединяющего разные однородные области изображения в один кластер, а также зависимость от первоначально задаваемых центров кластеров.

На рис. 1 и 2 приведены результаты сравнения GPFCM с КМ и FCM. Предложенный метод GPFCM демонстрирует гораздо большую устойчивость к шуму. Он может быть использован не только для кластеризации изображений, но и как вполне адекватный фильтр.

Заключение. Разработана новая версия метода нечеткой кластеризации k -средних с аддитивной штрафной функцией гиббсовского типа. Показано, что использование метода GPFCM дает более точные кластерные представления спутниковых изображений, чем известные методы КМ и FCM. Результаты экспериментов продемонстрировали значительное превосходство GPFCM над КМ и FCM при обработке зашумленных или искаженных изображений. Настройка параметров локальной гладкости с помощью аддитивной штрафной функции гиббсовского типа позволяет контролировать свойства кластеров для получения адекватных результатов. При нулевых значениях параметров метод GPFCM превращается в метод FCM. В будущем планируется провести более детальное исследование свойств GPFCM с применением различных значений показателей функций принадлежности и типов связей между пикселями изображений.

Список использованной литературы

1. *Steinhaus, H.* Sur la division des corps materiels en parties / H. Steinhaus // Bull. Acad. Polon. – 1956. – Vol. 4 (12). – P. 801–804.
2. *Lloyd, S.* Least square quantization in PCM's / S. Lloyd. – Bell Telephone Laboratories Paper, 1957.
3. *Bezdek, J. C.* Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – [S. l.]: Kluwer Academic Publishers Norwell, 1981.
4. *Гонсалес, Р.* Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М.: Техносфера, 2005.
5. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition* / F. Höppner [et al.]. – New York: John Wiley & Sons, 1999.
6. *MacQueen, J. B.* Methods for classification and Analysis of Multivariate Observations / J. B. MacQueen // Proc. 5th Berkeley Symp. on Math. Statistics and Probability. – Berkeley, 1967. – P. 281–297.
7. *Cenitha, C. H.* Classification of satellite images using new fuzzy cluster centroid for unsupervised classification algorithm // C. H. Cenitha, Dr. K. Vani // Proc. IEEE Conf. on Inform. and Commun. Technologies ICT, 2013. – [S. l.], 2013. – P. 203–207.
8. *Залесский, Б. А.* Метод нечеткой кластеризации k -средних со сглаживающей штрафной функцией / Б. А. Залесский // Информатика. – 2014. – № 43. – С. 14–20.

Поступила в редакцию 29.10.2015