NONPARAMETRIC CLUSTERING ALGORITHM FOR IMAGE SEGMENTATION COMBINING GRID-BASED APPROACH AND MEAN-SHIFT PROCEDURE

Sergey A. Rylov

Institute of Computational Technologies SB RAS, Novosibirsk, Russia

Abstract

Nonparametric clustering is attractive because of its ability to discover arbitrary shaped clusters. Mean-shift is a well-known mode-seeking procedure that is capable of producing accurate results. However, it has high time complexity. On the other hand, grid-based methods are computationally efficient, but their accuracy is limited by the grid structure. A novel clustering algorithm that combines grid-based HCA algorithm with mean-shift procedure is proposed. Experiments show significant improvement of clustering accuracy over HCA and high computational efficiency.

Keywords: grid-based approach, nonparametric clustering, mean-shift, image segmentation

НЕПАРАМЕТРИЧЕСКИЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ ДЛЯ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ НА ОСНОВЕ КОМБИНАЦИИ СЕТОЧНОГО ПОДХОДА И ПРОЦЕДУРЫ СРЕДНЕГО СДВИГА

Рылов С.А.

Институт вычислительных технологий СО РАН, Новосибирск

Непараметрические алгоритмы кластеризации позволяют выделять кластеры сложной формы. Однако вычислительная трудоемкость распространенных методов затрудняет их применение к обработке изображений. Сеточные алгоритмы кластеризации являются вычислительно эффективными, но при этом точность выделения кластеров зависит от сеточной структуры. В работе предлагается новый алгоритм кластеризации, основанный на сеточном алгоритме HCA и использовании процедуры среднего сдвига для уточнения границ кластеров. Экспериментальные исследования показывают увеличение точности кластеризации и демонстрируют высокую вычислительную эффективность алгоритма.

Ключевые слова: сеточный алгоритм кластеризации, непараметрический подход, процедура среднего сдвига, сегментация изображений.

Введение. При решении целого ряда прикладных задач возникает необходимость кластеризации больших массивов данных. Например, использование алгоритмов кластеризации является одним из наиболее распространенных подходов к сегментации мультиспектральных спутниковых изображений [1]. При этом априорные сведения о вероятностных характеристиках классов, а также обучающие выборки, как правило, отсутствуют. В этих условиях наиболее подходящими являются непараметрические алгоритмы, которые не требуют жестких предположений о виде функции плотности распределения и позволяют выделять кластеры сложной формы [2,3]. Однако они не получили распространения на практике из-за свойственной им высокой вычислительной трудоемкости. Применение сеточного подхода, при котором пространство признаков разделяется на конечное число ячеек, позволяет добиться высокой вычислительной эффективности, но при этом точность выделения кластеров сильно зависит от сеточной структуры [4].

В данной работе предлагается новый непараметрический алгоритм кластеризации, основанный на сеточном алгоритме HCA и использовании процедуры «среднего сдвига» (Mean-shift) для уточнения границ кластеров. Комбинация этих подходов позволяет получить одновременно высокую точность разделения кластеров и высокое быстродействие.

Краткое описание сеточного алгоритма кластеризации HCA. Предлагаемый алгоритм основывается на иерархическом сеточном алгоритме кластеризации HCA [5], который обладает высокой вычислительной эффективностью и способен выделять кластеры сложной формы. Далее приведено его краткое описание.

Пусть множество объектов X состоит из векторов, лежащих в пространстве признаков R^d: $X = \{x_i = (x_i^1,...,x_i^d) \in R^d, i = \overline{1,N}\}$, и ограниченных гиперпараллелепипедом $\Omega = [l^1, r^1] \times ... \times [l^d, r^d]$: $l^j = \min_{x_i \in X} x_i^j$, $r^j = \max_{x_i \in X} x_i^j$. *Сеточная структура* определяется как разбиение пространства признаков на *клетки* гиперплоскостями: $x^j = (r^j - l^j) \cdot i / m + l^j$, $i = 0,...,m$, где m – число разбиений Ω по каждой размерности. Множество клеток, смежных с B, обозначается через A_B. *Плотность* D_B клетки B определяется как число элементов множества X, попавших в клетку B.

Непустая клетка B_i *непосредственно связна* с непустой клеткой B_j $(B_i \to B_j)$, если B_j – максимальная по номеру клетка, удовлетворяющая условиям: $B_j = \arg\max_{B_k \in A_{B_i}} D_{B_k}$ и $D_{B_j} \geq D_{B_i}$. Непустые смежные клетки B_i и B_j непосредственно связны $(B_i \leftrightarrow B_j)$, если

$B_i \to B_j$ или $B_j \to B_i$. Непустые клетки B_i и B_j *связны* $(B_i \sim B_j)$, если существуют $k_1,...,k_l$ такие, что $k_1 = i,\ k_l = j$ и для всех $p = 1,...,l-1$ выполнено $B_{k_p} \leftrightarrow B_{k_{p+1}}$. Введение отношения связности порождает разбиение множества непустых клеток на компоненты связности $\{G_1,\ldots,G_S\}$. Под *компонентой связности* понимается максимальное множество попарно связных клеток. *Представителем компоненты связности* G называется максимальная по номеру клетка $Y(G)$, удовлетворяющую условию $Y(G) = \arg\max_{B \in G} D_B$.

Выделенные компоненты связности соответствуют одномодовым кластерам, а их представители – модам плотности этих кластеров. Далее, для построения иерархии между компонентами вводится специальная метрика.

Расстояние h_{ij} между смежными компонентами G_i и G_j определяется по формуле

$$h_{ij} = \min_{P_{ij} \in \Re_{ij}} \left[1 - \min_{B_{k_t} \in P_{ij}} D_{B_{k_t}} / \min(D_{Y_i}, D_{Y_j}) \right],$$

где $\Re_{ij} = \{P_{ij}\}$ – множество всех цепочек, связывающих представителей компонент связности, $P_{ij} = \left\langle Y(G_i) = B_{k_1},...,B_{k_t},B_{k_{t+1}},...,B_{k_l} = Y(G_j) \right\rangle$ таких, что для всех $t = 1,...,l-1$: 1) $B_{k_t} \in G_i \cup G_j$; 2) B_{k_t}, $B_{k_{t+1}}$ – смежные клетки.

После формирования матрицы расстояний между смежными компонентами $\{h_{ij}\}$, к ней применяется алгоритм построения дендрограммы методом ближайшего соседа (SLINK). В результате получается иерархическая структура на множестве компонент связности.

Алгоритм HCA при низких вычислительных затратах позволяет выделять сложные многомодовые кластеры, а также получать иерархическое представление результатов. Однако точность разделения кластеров зависит от сеточной структуры, что может приводить к ошибкам, особенно при неудачном выборе параметра масштаба сетки m.

Краткое описание процедуры «среднего сдвига». Для непараметрической оценки плотности распределения данных одной из наиболее широко используемых является оценка Розенблатта–Парзена [6]. Плотность оценивается как суммарное влияние элементов выборки, при этом вклад каждого элемента описывается колоколообразной функцией (*ядром*) $K(x)$, зависящей от расстояния до этого элемента. Формула для вычисления оценки плотности $f(x)$ с параметром сглаживания h в произвольной точке x имеет вид:

$$\hat{f}_h(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h} \right).$$

В качестве $K(x)$ можно использовать классическое ядро Гаусса:

$$K_G\left(\frac{x - x_i}{h} \right) = \exp\left(-\frac{\|x - x_i\|}{2h^2} \right).$$

Однако на практике, в целях сокращения вычислительных затрат используются ограниченные ядра, такие как например ядро Епанечникова:

$$K_{Ep}\left(\frac{x - x_i}{h} \right) = \left(1 - \frac{\|x - x_i\|^2}{h^2} \right) \cdot I\left(\|x - x_i\| \le h^2 \right),$$ где $I(x)$ – индикаторная функция.

В данном подходе кластеры соответствуют локальным максимумам функции оценки плотности (модам). А элементы данных относятся к кластерам с помощью процедуры «среднего сдвига» (Mean-shift) [3,7], сходящейся по градиенту к соответствующему локальному

максимуму. Итеративная процедура, начиная свою работу с точки x_0, последовательно перемещается в точку сдвига $x_{k+1} = m(x_k)$ вплоть до сходимости, где:

$$m(x) = \frac{\sum_{i=1}^{N} x_i \cdot K(x - x_i)}{\sum_{i=1}^{N} K(x - x_i)}.$$

Вектор $(m(x) - x)$ называется вектором «среднего сдвига» и его направление совпадает с направлением максимального роста плотности в точке x.

Алгоритмы кластеризации на основе использования процедуры среднего сдвига позволяют получать качественные разбиения, однако основной проблемой для использования этого подхода при обработке изображений является высокая вычислительная сложность [3,6-8].

Новый алгоритм кластеризации на основе комбинации сеточного подхода и процедуры среднего сдвига. Предлагаемый алгоритм основан на использовании сеточного алгоритма кластеризации HCA, при этом к элементам граничных клеток применяется процедура «среднего сдвига», в результате чего происходит уточнение границ получаемых кластеров. Далее приведено описание разработанного алгоритма HCA-MS.

На первом этапе происходит выполнение представленного выше алгоритма HCA с заданным параметром сетки *m*. В результате на сеточной структуре выделяются компоненты связности, на множестве которых строится иерархия.

На втором этапе осуществляется индексирование элементов данных по клеткам, позволяющее иметь быстрый доступ к списку элементов произвольной клетки.

На третьем этапе рассматриваются непустые клетки, находящиеся на границах компонент связности (т.е. такие, для которых существует смежная клетка, принадлежащая другой компоненте). К каждому элементу такой клетки применяется процедура «среднего сдвига» с ограниченным ядром и параметром *h*, равным ширине клетки в сеточной структуре. При этом для поиска элементов в радиусе *h* достаточно перебрать элементы только из смежных клеток. Данный процесс останавливается, если рассматриваемый элемент перемещается в другую непустую клетку. Если новая клетка принадлежит другой компоненте связности, то элемент перемещается в эту компоненту. Максимальное число итераций ограничивается параметром.

После коррекции границ между компонентами связности, соответственно оказываются скорректированными и границы получаемых кластеров на всех уровнях иерархии.

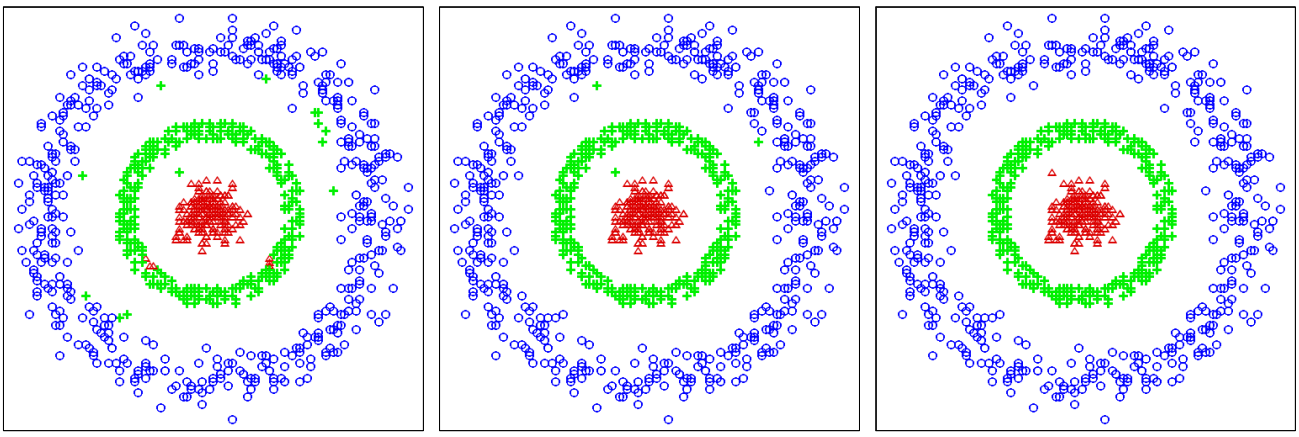
С точки зрения реализации предложенного алгоритма можно отметить следующие детали. 1) Максимальное число итераций процедуры «среднего сдвига» было ограничено тремя. Данное значение было выбрано эмпирически на основе экспериментальных исследований. 2) В процессе работы алгоритма формируется и используется таблица весов (число повторений элементов с совпадающими значениями векторов признаков). Для изображений характерна высокая частота повторяемости векторов спектральных яркостей, поэтому использование весов существенно сокращает вычислительные затраты. При этом значения всех признаков приводятся к диапазону $[0, 255]$ с 256-ю уровнями квантования. 3) Обработку граничных клеток можно проводить независимо друг от друга, поэтому третий этап был распараллелен для выполнения на ядрах центрального процессора.

Экспериментальные исследования. В данном разделе представлены результаты экспериментальных исследований предложенного алгоритма HCA-MS на модельных данных и изображениях. Приведено сравнение времени работы и точности кластеризации алгоритмов HCA-MS, HCA и Mean-shift.

Указанные алгоритмы были реализованы на языке программирования Java. Вычисления проводились на ПЭВМ с процессором Intel Core i5, 3.5 ГГц (4 ядра). При реализации алгоритма Mean-shift также использовалось индексирование элементов данных по клеткам, таблица весов и распараллеливание. Максимальное число итераций было ограничено десятью.

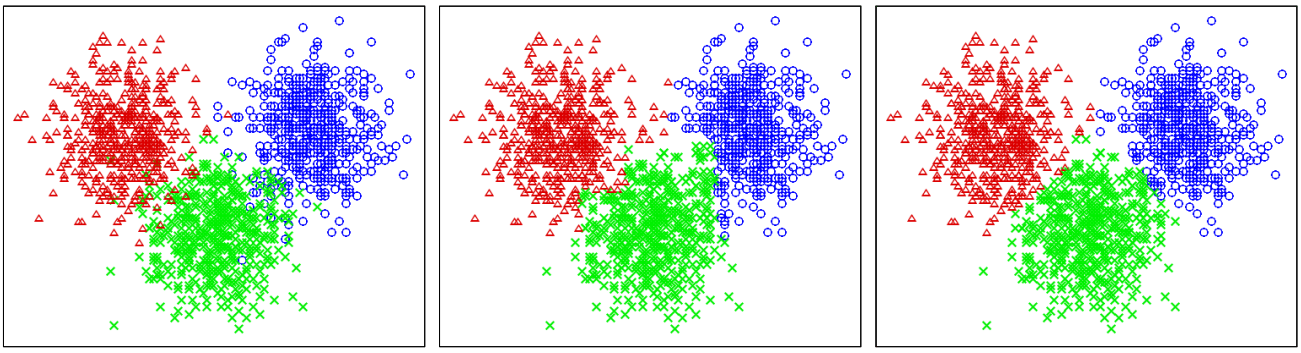
Эксперимент 1. Использовались двумерные модельные данные, в которых один кластер, описываемый нормальным распределением, окружен двумя кластерами в форме колец (рис. 1,в) [9]. Алгоритм кластеризации Mean-shift не способен выделить многомодовые кластеры в форме колец. В свою очередь алгоритм HCA позволяет успешно выделить все три кластера, однако допускает ошибки при неудачном выборе параметра сетки: при m=30 точность кластеризации составляет 98.21% (рис. 1,а); при m=42 допущена ошибка в 3 точках (рис. 1,б); а при m=46 получена точность 100% (рис. 1,в). Использование предложенного алгоритма позволяет скорректировать ошибки, вызванные сеточным эффектом: в результате работы HCA-MS во всех трех случаях получается безошибочный результат (рис. 1,в).

Эксперимент 2. На рис. 2,а представлена модель с тремя сильно пересекающимися нормально-распределенными кластерами [9]. На рис. 2,б представлен результат кластеризации этой модели алгоритмом HCA при m=20, точность составила 94.93%. В данном случае точность можно повысить, используя более мелкую сетку: например, при m=40 точность – 96%. Однако, измельчение сетки может быть недопустимо при выделении кластеров сложной структуры. В тоже время, алгоритм HCA-MS при m=20 демонстрирует точность 96.4% (рис. 2,в). Алгоритм Mean-shift достигает точности 96.33% при h=27.



а б в

Рис. 1. Результаты кластеризации модельных данных алгоритмом HCA при m=30 (а), m=42 (б), m=46 (в) и результат кластеризации HCA-MS при тех же параметрах (в).



а б в

Рис. 2. Результаты кластеризации модельных данных (а) алгоритмом HCA (б) и алгоритмом HCA-MS при m=20 (в).

Эксперимент 3. В таблице ниже приведено сравнение времени работы алгоритмов HCA-MS, HCA и Mean-shift на изображениях разного размера. В эксперименте использовались как цветные фотоизображения, так и мультиспектральные снимки, полученные со спутника WorldView-2 (кластеризация проводилась по 4-м каналам: 1, 3, 5, 7). Обрабатывавшиеся изображения [10] представлены на рис. 3 (в том же порядке, как в таблице). Параметры алгоритмов (m и h) были выбраны исходя из желания получить адекватные и схожие результаты

сегментации. Результаты эксперимента показывают, что предложенный алгоритм характеризуется существенно меньшей вычислительной трудоемкостью по сравнению с алгоритмом Mean-shift и позволяет обрабатывать цветные изображения большого размера за несколько секунд, а мультиспектральные – за несколько минут.

Сравнение времени работы алгоритмов HCA, HCA-MS и Mean-shift (в секундах).

Размер изображения (млн пикселей)	**Число каналов**	**HCA**		**HCA-MS**		**Mean-shift**	
		m=25	**m=32**	**m=25**	**m=32**	**h=20**	**h=25**
1	3	0.05	0.05	0.7	0.3	52	58
5	3	0.1	0.11	1.2	0.7	67	90
14	3	0.2	0.2	7.3	3.7	388	563
4	4	0.2	0.3	71	27	4138	6009
12	4	0.4	0.5	458	350	62388	97121

Рис. 3. Тестовые изображения: 3 цветные фотографии и 2 мультиспектральных спутниковых снимка.

Заключение. В работе предложен алгоритм кластеризации, основанный на сеточном иерархическом алгоритме HCA и использовании процедуры «среднего сдвига» для уточнения границ кластеров. Результаты экспериментов на модельных данных показывают, что предложенный алгоритм позволяет исправлять ошибки, обусловленные сеточным эффектом, тем самым повышая точность кластеризации и упрощая процедуру настройки параметра сетки *m*. Вычислительная эффективность алгоритма HCA-MS позволяет применять его к мультиспектральным спутниковым изображениям большого размера (состоящим из миллионов пикселей).

В дальнейшем планируется проверить, насколько значимый эффект оказывает уточнение границ кластеров на качество результатов сегментации спутниковых изображений, а также реализовать предложенный алгоритм для выполнения на графических процессорах.

ЛИТЕРАТУРА

[1] Xie Y., Sha Z., Yu M. Remote sensing imagery in vegetation mapping: a review // Journal of plant ecology. 2008. Vol. 1. No. 1. P. 9-23.

[2] Sarmah S., Bhattacharyya D.K. A grid-density based technique for finding clusters in satellite image // Pattern Recognition Letters. 2012. Vol. 33. No. 5. P. 589-604.

[3] Пестунов И.А., Синявский Ю.Н. Анализ и синтез сигналов и изображений непараметрический алгоритм кластеризации данных дистанционного зондирования на основе grid-подхода // Автометрия. 2006. Т. 42. №. 2. С. 90-99.

[4] Krstinic D., Skelin A.K., Slapnicar I. Fast two-step histogram-based image segmentation // Image Processing, IET. 2011. Vol. 5. No. 1. P. 63-72.

[5] Пестунов И.А., Рылов С.А., Бериков В.Б. Иерархические алгоритмы кластеризации для сегментации мультиспектральных изображений // Автометрия. 2015. Т. 51. № 4. С. 12-22.

[6] Пестунов И.А., Бериков В.Б., Синявский Ю.Н. Сегментация многоспектральных изображений на основе ансамбля непараметрических алгоритмов кластеризации // Вестн. Сиб. Гос. аэрокосмического ун-та им. академика М.Ф. Решетнева. 2010. № 5(31). С. 56-64.

[7] Cheng Y. Mean shift, mode seeking, and clustering // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995. Vol. 17. No. 8. P. 790-799.

[8] Freedman D., Kisilev P. Fast mean shift by compact density representation // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2009. P. 1818-1825.

[9] Рылов С.А. Модельные данные для кластеризации [Электронный ресурс]. URL: **https://drive.google.com/open?id=0ByK9GtU5ExExRnZwdFNmRHRWdFk** (дата обращения 30.06.2017).
[10] Тестовые изображения для кластеризации [Электронный ресурс]. URL: **https://drive.google.com/open?id=0ByK9GtU5ExExWXpGRjU5WVFHcDg** (дата обращения 30.06.2017).