Белоглазов Денис Александрович
Бублей Сергей Евгеньевич
Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: fin_val_iv@tsure.ru.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 88634371773.

Beloglazov Denis Aleksandrovich
Bubley Sergey Evgen'evich
Taganrog Institute of Technology – Federal State-Owned Educational Establishment of Higher Vocational Education "Southern Federal University".
E-mail: fin_val_iv@tsure.ru.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: +78634371773.

УДК 004.932.2

Е.А. Вершовский

МЕТОД КОНТРОЛЯ КАЧЕСТВА КЛАСТЕРИЗАЦИИ МУЛЬТИСПЕКТРАЛЬНОГО ИЗОБРАЖЕНИЯ

Рассматривается задача сравнения результатов кластеризации мультиспектральных изображений. Предлагается разработанный автором для её решения метод контроля качества кластеризации. Приведены результаты сравнения метода контроля качества кластеризации с известным методом матрицы ошибок. Рассматривается применение этой задачи на практике.

Данные дистанционного зондирования; матрица ошибок; матрица соответствия; метод контроля; качество кластеризации.

E.A. Vershovsky

QUALITY CONTROL METHOD FOR REMOTE SENSING DATA CLUSTERING

Problem is considered about remote sensing data clustering results comparing. It is offered designed by author for her decisions method to clustering quality control. The broughted results of the comparison of the clustering quality control method with well known error matrix method. It is considered using of this problem in practice.

Remote sensing data; error matrix; accuracy matrix; clustering quality; control method.

При проведении автоматического контроля качества кластеризации мультиспектрального снимка возникает ряд сложностей. На сегодняшний день для задачи кластеризации мультиспектральных данных не существует, так называемых, «бенчмарок» – общепризнанных тестовых наборов данных и заданий, с помощью которых можно оценить процент правильности кластеризации того или иного используемого алгоритма [1]. Под правильностью, в данном случае, понимается общий процент совпадения всех точек каждого кластера всем заранее известным точкам соответствующего класса поверхности, полученным методом полевых исследований или натурных наблюдений. В качестве примера бенчмарки для общей задачи кластеризации можно упомянуть *Fisher's Iris data set* [2,3]. Однако бенчмарки для общей задачи кластеризации не могут использоваться для оценки качества кластеризации мультиспектральных данных.

Создание тестового набора для автоматического контроля качества кластеризации мультиспектрального снимка на основе данных дистанционного зондирования и наземного обследования территории экономически невыгодно в силу больших финансовых затрат, связанных с проведением различных съемок на огромной территории с рельефом различной сложности, не всегда доступным для полевого измерения и составления эталонной тематической карты местности.

Ситуация контроля качества усложняется в случае, когда для кластеризуемого мультиспектрального снимка просто не существует эталонной тематической карты, с которой можно было бы провести сравнение результатов кластеризации, и таких случаев – подавляющее большинство.

Решение сложившейся проблемной ситуации возможно в двух направлениях:

- создание тестового набора данных для задачи кластеризации мультиспектрального снимка;
- определение метода оценки сравнения кластерных карт.

В качестве метода оценки сравнения кластерных карт предлагается использовать модификацию матрицы ошибок.

Матрица ошибок представляет собой инструмент, использующий кросс-табуляцию для анализа того, как соотносятся значения совпадающих классов, полученные из различных источников [4]. Матрица ошибок (табл. 1) предполагает предопределенность классов в обоих наборах данных (классы A-Z) и основывается на их совпадениях (главная диагональ).

Таблица 1

Матрица ошибок

		Результат кластеризации					
		A	B	C	…	Z	∑
Эталонная карта кластеров	A	n_{AA}	n_{AB}	n_{AC}	n_{Ai}	n_{AZ}	n_{A-}
	B	n_{BA}	n_{BB}	n_{BC}	n_{Bi}	n_{BZ}	n_{B-}
	C	n_{CA}	n_{CB}	n_{CC}	n_{Ci}	n_{CZ}	n_{C-}
	…	n_{iA}	n_{iB}	n_{iC}	n_{ii}	n_{iZ}	n_{i-}
	Z	n_{ZA}	n_{ZB}	n_{ZC}	n_{Zi}	n_{ZZ}	n_{Z-}
	∑	n_{-A}	n_{-B}	n_{-C}	n_{-i}	n_{-Z}	N

В ячейках таблицы находится количество точек, располагающихся одновременно в классах столбца и строки. На главной диагонали находится количество совпавших точек для каждого класса. Другими словами, точка, принадлежащая классу C эталонной карты и классу B в результирующем наборе, прибавляется к значению n_{CB} ячейки. Остальные значения матрицы ошибок вычисляются по следующим формулам:

$$n_{-A} = n_{AA} + n_{BA} + n_{CA} + \ldots + n_{iA} + \ldots + n_{ZA};$$
$$n_{A-} = n_{AA} + n_{AB} + n_{AC} + \ldots + n_{Ai} + \ldots + n_{AZ}; \quad (1)$$
$$N = \sum n_{-i} = \sum n_{i-}.$$

Внедиагональные элементы показывает случаи несовпадения между рассчитанными и реальными классами (ошибки классификации).

Сумма значений диагональных элементов показывает общее количество правильно кластеризованных пикселей, а отношение этого количества к N – общему количеству пикселей в матрице называется общей точностью (*Overall Accuracy, OAcc*) классификации и выражается в процентах:

$$OAcc = \frac{n_{aa} + n_{bb} + n_{cc} + n_{dd} + ... + n_{ii} + ... + n_{zz}}{N}. \quad (2)$$

Для определения точности определенного рассчитанного класса необходимо разделить количество правильно классифицированных пикселей этого класса на общее количество пикселей в этом классе согласно проверочным данным. Этот показатель также называют «точностью производителя» (*Producer's Accuracy, PAcc*). Для класса А:

$$PAcc = \frac{n_{AA}}{n_{-A}} \quad . \quad (3)$$

С показателем точности производителя связано понятие ошибок омиссии (*Ommission error, OErr*). Данный показатель иллюстрирует процент пропуска пикселей, которые на самом деле (согласно проверочному набору данных) принадлежат определенному классу (кластеру), однако в результирующем наборе относятся к другому кластеру. Ошибка омиссии связана с точностью производителя следующим выражением:

$$OErr_A = 1 - PAcc_A \ . \quad (4)$$

Аналогичный точности производителя показатель может быть вычислен для реального класса, если разделить количество правильно кластеризованных пикселей класса на общее количество пикселей в этом классе согласно проверяемым данным. Показатель «точность пользователя» (*User's Accuracy, UAcc*) показывает пользователю классификации насколько вероятно, что данный класс совпадает с результатами классификации. Для класса А:

$$UAcc = \frac{n_{AA}}{n_{A-}}. \quad (5)$$

С показателем точности пользователя связано понятие ошибок комиссии (*Commission error, CoErr*). Этот показатель противоположен по смыслу ошибке омиссии, так как иллюстрирует количество пикселей, которые в результате кластеризации были отнесены к заданному классу, но согласно проверочному набору данных являются элементами других кластеров. Ошибка комиссии связана с точностью пользователя выражением:

$$CoErr_A = 1 - UAcc_A \ . \quad (6)$$

У приведенного метода использования матрицы ошибок есть два существенных недостатка [5], применительно к задаче кластеризации мультиспектральных снимков:

1. Необходимость взаимно-однозначного соответствия классов в разных наборах данных. Иными словами, после проведения кластеризации класс А в «проверяемом наборе» должен соответствовать классу А в результирующем наборе. Применение автоматической кластеризации не включает сопоставление получившихся классов с целью их именования согласно именованию классов в проверяемом наборе.
2. Необходимость иметь в наличии проверочный набор с заранее известной кластеризационной картой.

Эти недостатки предлагается разрешить следующим образом: необходимо унифицировать метод матрицы ошибок для общего случая, когда не существует проверочного набора с заранее определенными классами с учетом неопределенности взаимного соответствия классов в получившихся наборах. В качестве прове-

рочного набора предлагается использовать результат автоматической кластеризации одним из классических алгоритмов (*k-means, ISODATA*), при условии, что итоговое количество полученных классов в обоих результатах совпадает. При этом матрица ошибок преобразовывается в матрицу соответствия, аналогичную, однако отличную от классической матрицы корреляции. Подход позволяет проводить сравнение результатов работы различных алгоритмов кластеризации, сравнение которых по другим критериям (вычислительная сложность, скорость выполнения, требуемое для сходимости количество итераций и т.п.) затруднено ввиду различий в реализации и используемых фундаментальных основаниях каждого из алгоритмов (например, Fuzzy c-means и ISODATA).

Рассмотрим матрицу соответствия (табл. 2). В ней *A1, A2,..., An* – классы, полученные в результате кластеризации снимка алгоритмом A, *B1, B2…Bn* – классы, полученные в результате кластеризации снимка алгоритмом B. В принципе, это может быть один и тот же алгоритм, но с разными инициализационными параметрами (способ инициализации центров кластеров, количество итераций и т.п.). Изначально неизвестно, какому из классов *A1, A2,..., An* соответствует класс из *B1, B2,..., Bn*. Существует лишь требование в виде равенства общего количества кластеров в результате работы обоих алгоритмов. В ячейках табл. 2 находится количество точек, располагающихся одновременно в классах столбца и строки.

Таблица 2

Матрица соответствия

		Результат кластеризации алгоритмом B					
		B1	B2	B3	…	Bn	$\sum_{Aj}$
Результат кластеризации алгоритмом A	A1	c_{A1B1}	c_{A1B2}	c_{A1B3}	c_{A1Bi}	c_{A1Bn}	c_{A1-}
	A2	c_{A2B1}	c_{A2B2}	c_{A2B3}	c_{A2Bi}	c_{A2Bn}	c_{A2-}
	A3	c_{A3B1}	c_{A3B2}	c_{A3B3}	c_{A3Bi}	c_{A3Bn}	c_{A3-}
	…	c_{AjB1}	c_{AjB2}	c_{AjB3}	c_{AjBi}	c_{AjBn}	c_{Aj-}
	An	c_{AnB1}	c_{AnB2}	c_{AnB3}	c_{AnBi}	c_{AnBn}	c_{An-}
	$\sum_{Bi}$	c_{-B1}	c_{-B2}	c_{-B3}	c_{-i}	c_{-Bn}	N

Остальные значения матрицы соответствия в табл. 2 вычисляются по следующим формулам:

$$c_{-Bi} = \sum c_{AjBi}, j=1...n;$$
$$c_{Aj-} = \sum c_{AjBi}, i=1...n; \quad (7)$$
$$N = \sum c_{-Bi} = \sum c_{Aj-}, i=1...n, j=1...n.$$

Тогда, соответствие классов определяется следующим образом:

1. Устанавливаем $j=1$.
2. Ищем максимальный элемент в строке матрицы $c_{AkBm} = \max(c_{AjBm})$.
3. Проверяем, является ли найденный элемент также максимальным и в столбце. Если $c_{AkBm} = \max(c_{AkBi})$, тогда класс Ak соответствует классу Bm. Исключаем k-ую строку и m-ый столбец из дальнейшего рассмотрения. В противном случае увеличиваем j.
4. Если $j=n$, завершаем процедуру, иначе переходим к шагу 1.

Как видно из алгоритма, максимальные перекрестные элементы по столбцам и строкам не обязательно располагаются на главной диагонали.

Обозначим через d_{AkBm} значение ячейки k-ой строки m-ого столбца, которое устанавливает соответствие класса Ak классу Bm и является максимальным в указанных строке и столбце.

Тогда сумма значений d_{AkBm} показывает общее количество одинаково кластеризованных пикселей, а отношение этого количества к N – общему количеству пикселей в матрице назовем общим соответствием (*Overall Accordance, OAcc*) кластеризации и выражается в процентах:

$$OAcc = \frac{\sum_{i,j=1}^{n} d_{AiBj}}{N} \ . \tag{8}$$

В предлагаемом подходе нет понятия «эталонных» данных, с которыми можно сверять правильность кластеризации. Поэтому вместо аналогичных показателей точности производителя и точности пользователя введем показатель точности соответствия для каждого алгоритма (*Accordance Accuracy, AccAcc*), который будет отображать отношение количества пикселей, лежащих на пересечении k-ой строки и m-ого столбца, к сумме значений по строке или по столбцу, в зависимости от того, оценка какого алгоритма необходима.

Показатель точности соответствия в рамках алгоритма A для класса *Ak*, которому соответствует класс *Bm* по результатам работы алгоритма B, определится следующим образом:

$$AccAcc_{Ak} = \frac{d_{AkBm}}{c_{Ak-}} \ . \tag{9}$$

Аналогично рассчитывается показатель точности соответствия в рамках алгоритма B для класса *Bm*, которому соответствует класс *Ak*:

$$AccAcc_{Bm} = \frac{d_{AkBm}}{c_{-Bm}} \ . \tag{10}$$

Ошибкой соответствия (*Accordance Error, AccErr*) класса *Ak* по результатам работы алгоритма A, которому соответствует класс *Bm*, полученный в результате работы алгоритма B, назовем выражение:

$$AccErr_{Ak} = 1 - AccAcc_{Ak} \ . \tag{11}$$

Выражение (11) иллюстрирует процент ошибочно отнесенных к другим классам пикселей. По ошибочностью понимается отнесение пикселей к классам, не являющимся ранее определенными как соответствующие (*AkBm*). Аналогично рассчитывается ошибка соответствия класса *Bm*:

$$AccErr_{Bm} = 1 - AccAcc_{Bm} \, . \tag{12}$$

Общий показатель точности соответствия (*Overall Accordance Accuracy, OvAccAcc*) рассчитывается следующим образом:

$$OvAccAcc_{A} = \frac{\sum_{i=1}^{n} AccAcc_{Ai}}{n} \ . \tag{13}$$

Аналогично, общий показатель ошибки соответствия (*Overall Accordance Error, OvAccErr*) для алгоритма может быть рассчитан следующим образом:

$$OvAccErr_{A} = \frac{\sum_{i=1}^{n} AccAcc_{Ai}}{n} \ . \tag{14}$$

Данный подход может быть усовершенствован введением этапа сравнения показателей среднеквадратичной ошибки с целью выявления наиболее правильного результата. В этом случае, отталкиваясь от результатов алгоритма с наименьшей среднеквадратичной ошибкой, необходимо обозначить его как эталонный, и выполнив описанные выше вычисления получить численное значение ухудшения/улучшения качества кластеризации, в сравнении с эталоном.

В общем случае для сравнения результатов кластеризации используют такие параметры как среднюю квадратичную ошибку, минимальное межкластерное расстояние, максимальное внутрикластерное расстояние и т.п. Однако это накладывает ограничения на сравниваемые алгоритмы. В них должны использоваться одни метрики и одна и та же фитнесс функция, а это является существенной проблемой для алгоритмов, различающихся внутренней архитектурой и используемым математическим аппаратом.

Рассмотрим пример, иллюстрирующий применение предлагаемого метода.

Имеются результаты работы алгоритма *K-means* с различными начальными параметрами (ограничение в 10 (A) и в 100 итераций (B)). В обоих случаях кластеризация производилась с параметром k = 3.

Составляем матрицу соответствия по описанному выше методу (табл. 3).

Серым цветом выделены элементы d_{AkBm}, устанавливающие соответствие класса *Ak* классу *Bm* и являющиеся максимальными в указанных строке и столбце.

Таблица 3

Матрица соответствия

		Результат кластеризации алгоритмом B					
		B1	B2	B3	$\sum_{Aj}$	$AccAcc_{Ai}$	$AccErr_{Ai}$
Результат кластеризации алгоритмом A	A1	11	15556	16383	31950	51	49
	A2	97044	197	0	97241	100	0
	A3	17159	105429	0	122588	86	14
	$\sum_{Bi}$	114214	121182	16383	251779		
	$AccAcc_{Bi}$	85	87	100			
	$AccErr_{Bi}$	15	13	0			

$$OvAcc = \frac{\sum_{i,j=1}^{n} d_{AiBj}}{N} = \frac{97044 + 105429 + 16383}{251779} = 87\,\%;$$

$$AccAcc_A = \frac{(85 + 87 + 100)}{3} = 91\,\%;\ AccErr_A == \frac{(15 + 13)}{3} = 9\,\%; \quad (15)$$

$$AccAcc_B = \frac{(51 + 100 + 86)}{3} = 79\ \%;\ AccErr_B == \frac{(49 + 14)}{3} = 21\,\%.$$

Таким образом, отношение совпавших пикселей в классах к общему числу пикселей изображения составляет 87 %. Алгоритм A обладает следующими показателями: средний процент кластеризованных пикселей к совпавшим равен 79 %, отклонение – 21 %, среднеквадратичная ошибка алгоритма – 45,7. Алгоритм B обладает следующими показателями: средний процент кластеризованных пикселей по отношению к совпавшим равен 91 %, отклонение – 9 %, среднеквадратичная ошибка алгоритма – 39,8.

Учитывая сравнение среднеквадратичных ошибок обоих результатов, можно определить насколько хуже алгоритм B справился с задачей автоматической кластеризации мультиспектрального снимка, по отношению к алгоритму A, не в относительных, а в абсолютных значениях. Для этого требуется выполнить следующие вычисления:

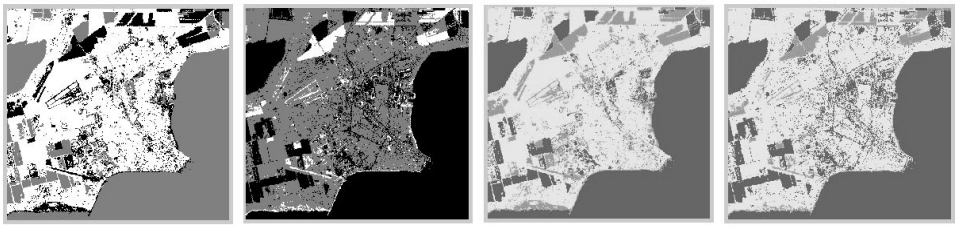
$$OvErr_{Bi} = \frac{\left|\sum A_j - \sum B_i\right|}{\sum B_i}, \quad (16)$$

где $j=k$, $i=m$, $d_{AkBm} = \max(c_{AjBm})$ и $d_{AkBm} = \max(c_{AkBi})$,

$$OvErr_B = \frac{\sum_{i=1}^{n} OvErr_{Bi}}{n}. \quad (17)$$

В данном примере, результаты работы алгоритма B хуже на 36 %.

Еще одной немаловажной стороной предложенного метода является установка соответствия между кластеризационными картами разных результатов кластеризации. При проведении неконтролируемой классификации изображений, начальные центры кластеров в большинстве алгоритмов инициализируются псевдослучайным образом, что существенно влияет на раскраску конкретных классов после кластеризации. Этим обстоятельством объясняется то, что в большинстве случаев, при опубликовании результатов работы алгоритмов кластеризации их карты не имеют общей цветовой схемы (рис. 1,а,б), что затрудняет визуальное восприятие результатов, и, тем более, их сравнительную оценку. Предлагаемый метод решает эту проблему, позволяя применять единую цветовую схему для различных результатов кластеризации за счет автоматического определения соответствия классов (рис. 1,в,г).



а б в г

Рис. 1. Результаты кластеризации: а, б – алгоритмов A и B без проведения соответствия классов, в, г – алгоритмов A и B после проведения соответствия

Предложенные показатели позволяют проводить оценку результатов алгоритмов кластеризации различных по своей архитектуре и вычислительной сложности алгоритмов в отсутствие эталонной кластеризационной карты, при кластеризации мультиспектральных данных дистанционного зондирования в случае совпадения итогового количества полученных классов. Метод прошел апробацию при экспериментальном исследовании роевого алгоритма кластеризации в сравнении с классическими алгоритмами кластеризации [6].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Chulhee L.,Landgrebe D.* Analyzing High Dimensional Multispectral Data // IEEE Transactions on Geoscience and Remote Sensing, 31. –1993. – № 4. – P. 792-800.
2. *МакКоннелл Д.* Основы современных алгоритмов. – М.: Техносфера, 2004. – 368 с.
3. *Дубровский С.А.* Прикладной многомерный статистический анализ. – М.: Финансы и статистика, 1982. – 216 с.
4. Матрица ошибок и расчет показателей точности тематических карт [Электронный ресурс] // Gis-Lab: Геоинформационные системы и Дистанционное зондирование Земли [сайт]. URL: http://gis-lab.info/qa/error-matrix.html (дата обращения 13.01.2010).
5. Glossary of remote sensing terms [Электронный ресурс] // The Canada Centre for Remote Sensing [сайт]. URL: http://www.ccrs.nrcan.gc.ca/glossary/index_e.php?id=3124 (дата обращения 13.01.2010).
6. *Вершовский Е.А.* Роевой алгоритм оптимизации в задаче кластеризации мультиспектрального снимка // Известия ЮФУ. Технические науки. – 2010. – № 5 (106). – С. 102-107.

Вершовский Евгений Алексеевич
Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: vershovsky@cbt.ru.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 88634371673.

Vershovsky Eugeniy Alekseevich
Taganrog Institute of Technology – Federal State-Owned Educational Establisment of Higher Vocational Education «Southern Federal University».
E-mail: vershovsky@cbt.ru.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: +78634371673.

УДК 681.518

Ю.В. Клунникова

МОДЕЛЬ ВЛИЯНИЯ ПАРАМЕТРОВ ТЕХНОЛОГИЧЕСКОГО ПРОЦЕССА ПОЛУЧЕНИЯ САПФИРА НА КАЧЕСТВО КРИСТАЛЛОВ

Представлена модель влияния параметров технологического процесса получения монокристаллов сапфира на качество кристаллов. Разработано математическое и информационное обеспечение процесса получения кристаллов сапфира, которое систематизирует большие информационные массивы данных и дает точную характеристику кристаллов.

Информационная система; технологический процесс получения монокристаллов сапфира; качество; оптимизация.

Y.V. Klunnikova

MODEL OF THE SAPPHIRE GROWTH TECHNOLOGICAL PROCESS PARAMETERS INFLUENCE ON CRYSTALS QUALITY

The model of sapphire growth technological process parameters influence on crystals quality is presented in this article. The software and data ware for sapphire crystals growth process are developed. It allows to systematize the large information volumes and to give the exact crystals characteristics.

Information system; sapphire production technological process; quality, optimization.

Технологическими особенностями получения монокристаллических структур сапфира являются длительность этих процессов (от нескольких дней до нескольких недель), высокие температуры процессов, зависимость качества монокристаллических структур от режимов выращивания. Такие технологические процессы очень трудоемки. Современное состояние средств их автоматизации предполагает использование интегрированных информационных сред с применением технологий представления знаний при создании автоматизированных информационных систем. Это открывает неограниченные возможности использования автоматизированных систем для процессов получения монокристаллических структур сапфира. Сложность использования этих систем состоит в неполноте математического описания технологических моделей процессов кристаллизации сапфира [1].

Целью данной работы является разработка модели влияния параметров технологического процесса получения сапфира на качество кристаллов. Информационные системы рассматриваются в качестве инструментария для реализации поставленной цели исследования.