ОСНОВАНИЯ МНОГОМЕРНОГО КЛАСТЕРНОГО АНАЛИЗА В ГЕОЛОГИИ

Ю. А. Ткачев

Институт геологии Коми НЦ УрО РАН, Сыктывкар
tkachev@geo.komisc.ru

Охарактеризована роль кластерного анализа в системе наук о классиологии, даны определения кластеров в текстовом варианте и с помощью логико-математических формул. Впервые дано сопоставление идей и методов кластерного анализа с теорией графов. Изложен алгоритм автоматического выделения многомерных кластеров объединением так называемых строковых кластеров, под которыми понимаются кластеры, объединяемые наличием в них общей точки под номером, равным номеру строки. Установлено, что графы кластеров являются наглядной и исчерпывающей формой представления результатов кластерного анализа.

Кроме того, рассмотрены специальные вопросы: выбор критического расстояния, кластограммы, преобразование исходных данных для кластерного анализа, трактовка плоских (двухмерных) изображений многомерного кластерного анализа.

Ключевые слова: классиология, классификация, кластерный анализ, критическое расстояние, алгоритм, программа для ЭВМ, теория графов.

MULTIVARIATE CLUSTER ANALYSIS IN GEOLOGY

Yu. A. Tkachev

Institute of Geology of Komi SC, UB RAS, Syktyvkar

The article discusses the role of cluster analysis in the sciences of classiology, defines clusters using text as well as logical and mathematical formulas. For the first time, the ideas and methods of cluster analysis are discussed side by side with the graph theory. The algorithm is defined for the automatic recognition of multivariate clusters by joining string clusters — clusters with have a common point (a row in the table of point-to-point distances).

Separately, the following special issues are discussed: the choice of critical path, clustergrams, transformation of input data for cluster analysis, and interpretation of plain (two-dimensional) images of multivariate cluster analysis.

Keywords: classiology, classification, cluster analysis, critical path, algorithm, computer program, graph.

Место кластерного анализа в классиологии

Кластерный анализ является одним из методов, применяемых для разделения множества объектов и процессов на виды, классы, группы, типы и т. д., то есть на подмножества [15, 16, 22]. В классиологии выделяют два типа задач: формальные (математические) и содержательные (предметные). Первые служат для установления самого факта существования классов, групп и др. и отвечают на вопрос, не являются ли они результатом случайной флуктуации значений свойств, характеризующих объекты или процессы [16]. Достигается это проверкой статистических гипотез о существенности или несущественности вероятностей реальности наблюдаемых групп. К этому типу задач относятся дисперсионный и факторный анализы, дискриминантный анализ [13, 20], теория графов, теория размытых множеств (fuzzi) и анализируемый в этой статье кластерный анализ [22]. Среди классиологической литературы выделяются объемные монографии (до 600 стр. и более — [15, 16]), написанные профессиональными философами, но не содержащие ни одной логико-математической формулы. Из этих монографий можно почерпнуть мысль лишь о необходимости классифицировать объекты, свойства и процессы, но не о том, как это сделать.

Кластерный анализ заключается в разделении на группы (кластеры, подмножества) множеств изучаемых точек-объектов. В геологии информация обычно привязана к определенным точкам поверхности

Земли или её недр. Традиционно она изображается в виде карт и планов, на которых геологические тела условно отображаются в виде точек с двухмерными или трехмерными координатами. Пространственное распределение этих точек подвергается кластерному, а затем предметному анализу с целью установления геологических закономерностей и прогноза месторождений полезных ископаемых [1, 3, 6].

Однако область применения кластерного анализа не ограничивается точками-объектами в одно-, двух- или трехмерном физическом пространстве. Огромное количество геологической информации касается состава и свойств проб, образцов и других объектов. Эти данные отображаются в табличной форме, а также в графической форме в виде точек в абстрактном многомерном пространстве, в пространстве признаков или свойств [22]. Размерность пространства D в этом случае равна числу изучаемых признаков, которых в геологии в большинстве случаев бывает намного больше трех.

Начала кластерного анализа

С давних пор подразделение точек на кластеры производилось визуально. Ввиду наглядности графических изображений визуальный обзор и сейчас представляет большую ценность. Он позволяет оценивать расположение групп точек, их число в группах, плотность расположения точек в кластерах по сравнению с плотностью их расположения в среднем и в межкластерном пространстве. В этом и только в этом заклю-

чался прежде кластерный анализ, который назовем визуальным. Он всегда осуществлялся геологами при рассмотрении точечных графиков, но специально не выделялся как самостоятельный метод.

Несмотря на несомненную пользу, его возможности крайне ограничены. Наглядными фактически были только одномерный и двухмерный случаи. Для придания наглядности в трехмерном пространстве, исполненном на плоскости (на бумаге, экране компьютера), приходится использовать специальные средства или методы (стереоскопы, стереочки, цвета точек), однако их результаты мало впечатляют. Многомерный случай в принципе не может быть наглядным из-за приспособленности мозга животных и человека к трехмерному физическому пространству, в котором он возник и эволюционировал. Но с помощью абстрактных математических методов можно решать некоторые задачи о свойствах размещения точек в многомерном пространстве [10, 14]. Введение в кластерный анализ понятия *расстояние* было важным шагом в его развитии — он стал *количественным*, основанным на строгих определениях и правилах.

Определение кластера

Естественными представляются следующие два определения кластера: 1) для любой точки кластера найдется хотя бы одна точка этого же кластера, удаленная от неё на расстояние, меньшее заданного критического, 2) все точки кластера удалены друг от друга на расстояние, меньшее заданного критического. По формулам математической логики первый вариант записывается как $\forall n \in G \vee \exists n_i (P(d(n, n_i) \in G < r_{кр}),$ т. е. в первом варианте каждая точка кластера соединена расстоянием, меньшим критического, хотя бы с одной точкой этого же кластера. Здесь $\forall$ — квантор общности, $\exists$ — квантор существования, G — имя кластера, n — любая (каждая) точка кластера, n_i — какая-нибудь одна точка кластера, $r_{кр}$ — критическое расстояние, $d(i, j)$ — расстояние между парой точек i, j . По второму определению для каждой пары точек кластера должно быть $d(i, j) < r_{кр}$ со всеми точками кластера, т. е. $\forall n_i \in G \vee \forall n_j \in G (P(d(i, j) < r_{кр}))$. Из определений очевидно, что они характеризуют крайние варианты кластеров. Под первое определение подпадают *кластеры-цепочки*, образованные одномерной последовательностью точек, соседние точки которых находятся на расстояниях, меньших критического, а соседние соседних — на больших критического. Такие кластеры называются неполносвязанными. Под второе определение подпадают совокупности точек, все пары в которых отстоят друг от друга на расстояние, меньше критического. Они называются полносвязанными.

Некоторые исследователи признают кластерами только те множества точек, в которых каждая точка удалена от каждой иной на расстояние, меньшее критического. Но необходимо иметь в виду, что многие точки могут быть связаны с другими не непосредственно, а через цепочку третьих точек. Отказать им в близости к остальным точкам кластера — значит игнорировать большую часть практических результатов кластерных анализов. Можно дать ещё одно определение кластера: любые две точки кластера можно соединить непрерывным маршрутом, что становится оче-

видным из теории графов (см. последний раздел статьи). Иначе простые цепи надо будет исключить из числа кластеров.

Строковые кластеры

Введём в кластерный анализ новое понятие — *строковые кластеры как промежуточные объекты процесса поиска кластеров в обычном их понимании* — *полных кластеров*. Строковый кластер — множество точек строки i , находящихся от точки i на расстояниях, меньших заданного критического. Однако друг от друга эти точки могут находиться на расстояниях, меньших критического. Расстояния между точками образуют симметричную матрицу размером $n \times n$, где n — число точек в изучаемой совокупности. Для нашей цели в таблицу можно ввести только те расстояния, которые меньше критического. Строго говоря, нам понадобятся не сами расстояния между точками, а лишь произвольные отметки тех клеток матрицы, расстояния между точками в которых меньше критического. Назовём эти клетки *отмеченными*. *Определение: Строковым кластером $Cl(i)$ строки i является совокупность точек с номером строки (m , т. е. i) и точек с номерами столбцов отмеченных клеток этой строки.*

Поскольку все точки в отмеченных клетках строки i удалены от точки i на расстояние, меньшее критического, то все они по первому определению удовлетворяют понятию кластера.

Постулат 1: каждая строка может не составлять строкового кластера (если в строке нет отмеченных клеток) или составить один и только один строковый кластер.

Постулат 2: если строковые кластеры строк i и j , $i \neq j$ не содержат ни одной общей точки, то они образуют различные строковые кластеры, т. е. $Cl(i) \neq Cl(j)$.

Постулат 3: если строковые кластеры строк i и j , $i \neq j$ имеют одну или более общих точек, то сумма строковых кластеров $Cl(i) + Cl(j)$ составляет один кластер.

Постулат 4: если кластеры $Cl(i) + Cl(j)$ могут составить один кластер и $Cl(i) + Cl(t)$ могут составить один кластер, то $Cl(i) + Cl(t)$ также могут составить один кластер.

Это означает, что свойство строковых кластеров составлять один кластер является транзитивным. Таким образом, составление кластера в обычном смысле слова осуществляется суммированием строковых кластеров по правилам суммирования *множеств* точек строковых кластеров.

Остаётся добавить, что строковые кластеры наиболее часто будут неполно связанными с номером точки, равной номеру строки, с которой связаны все остальные точки.

Алгоритм с комментариями выделения множества точек, образующих многомерный кластер с помощью строковых кластеров

1. Исходными данными для алгоритма является таблица координат точек размером $n \times k$, где n — число точек, k — размерность пространства, т. е. число свойств.

Таблица 1. Химический состав зювитов из Карского импактного кратера по данным микрозондового анализа, мас. % (исходные данные для кластерного анализа, Н. С. Ковальчук, 9 компонентов, 19 проб)

Table 1. The chemical composition of the zuvits from the Kara impacting crater according to microsond analysis, masses. % (initial data for cluster analysis, N. S. Kovalchuk, 9 components, 19 samples)

№ пробы / Sample No	Na ₂ O	MgO	Al ₂ O ₃	SiO ₂	SO ₃	K ₂ O	CaO	TiO ₂	Fe ₂ O ₃
1	4.75	4.50	18.79	59.99	0.00	2.90	3.67	1.75	3.66
2	5.48	2.31	20.68	60.21	0.00	2.63	4.47	1.21	3.00
3	5.11	3.11	20.30	61.24	0.00	2.96	4.12	0.47	2.68
4	5.04	2.18	19.96	61.93	0.00	4.71	3.37	0.82	1.97
5	0.80	14.62	8.75	49.32	0.00	0.29	1.61	0.00	24.63
6	0.84	10.85	5.59	72.18	0.00	0.59	1.04	0.00	8.91
7	2.54	1.55	17.29	69.86	0.70	4.16	1.88	0.85	1.16
8	1.05	3.14	14.34	72.12	0.00	3.09	1.68	0.81	3.77
9	0.00	7.13	19.12	57.57	0.00	3.83	3.58	0.00	8.77
10	0.73	5.56	19.30	61.52	0.00	4.34	2.02	0.00	6.54
11	0.74	4.18	16.86	68.04	0.00	4.22	1.81	0.95	3.20
12	0.80	3.17	14.34	73.52	0.00	3.10	1.49	0.00	3.58
13	0.56	7.52	15.87	57.07	0.50	3.72	2.86	1.30	10.58
14	4.34	5.21	11.32	61.37	1.79	2.39	9.02	0.54	4.02
15	4.96	2.38	10.94	71.30	0.97	1.48	5.13	0.00	2.84
16	3.38	4.60	9.47	62.63	0.37	2.58	10.10	0.82	6.05
17	14.28	0.00	22.89	61.73	0.00	0.35	0.75	0.00	0.00
18	7.87	1.58	18.15	63.28	1.92	4.34	2.86	0.00	0.00
19	5.53	3.60	14.24	49.38	11.79	3.04	9.11	0.98	2.32
среднее / average	3.62	4.59	15.70	62.86	0.95	2.88	3.71	0.55	5.14
дисперсия / dispersion	12.07	12.16	22.06	49.11	7.26	1.86	7.87	0.31	30.75
станд. откл. / standard deviation	3.47	3.49	4.70	7.01	2.69	1.36	2.81	0.55	5.55

Например, это может быть таблица результатов анализов n проб на k химических элементов (табл. 1).

2. Исследователь принимает решение, следует ли подвергать исходные данные предварительному преобразованию. Оно может заключаться в нормировании, т. е. делении данных на *среднее значение* m , или на *среднее квадратическое отклонение* σ , или на *коэффициент вариации* v . При этом исходные данные заменяются на нормированные. На практике часто последовательно используют несколько вариантов нормирования. В качестве нормирующих коэффициентов исследователь может использовать весовые коэффициенты, отражающие относительную важность каждого признака.

3. Для каждой пары точек (i, j) рассчитывается евклидово расстояние по формуле:

$$d(i, j) = \sqrt{\sum_{t=1}^k (x_{i,t} - x_{j,t})^2},$$

где k — размерность пространства, $x_{i,t}$ — координата номера t точки i (результат анализа пробы i на химический элемент t). Значения $d(i, j)$ образуют симметричную матрицу $n \times n$ с нулевой главной диагональю — матрицу евклидовых расстояний (табл. 2).

4. По матрице евклидовых расстояний вычисляются минимальное $\min(d(i, j))$ и максимальное $\max(d(i, j))$ расстояния.

5. Исследователь намечает число критических расстояний, для которых будет проводиться кластер-

ный анализ. Практика показала, что результаты кластерного анализа сильно зависят от критического расстояния r , поэтому применяют серию критических расстояний, которые в совокупности образуют динамическую картину кластеров в зависимости от r .

Пусть число критических расстояний выбрано равным nr . Тогда вся дальнейшая обработка будет вестись с критическими nr -расстояниями, увеличивающимися с шагом

$$\frac{\max(d(i, j)) - \min(d(i, j))}{nr}.$$

6. Для каждого критического расстояния составляется таблица строковых кластеров, как описано в предыдущем разделе, причем каждая строка евклидовых расстояний может образовать пустой строковый кластер, несколько строковых кластеров хотя бы с одной общей точкой, объединяемых в единый кластер, и несколько строковых кластеров, объединяемых в другой кластер (табл. 3).

Выбор критического расстояния

На результаты кластерного анализа решающее влияние оказывает величина критического расстояния между точками. Если критическое расстояние выбрано слишком малым, образуется много малоточечных кластеров; если оно велико, в пределе получится один кластер, охватывающий все точки. На результат кластерного анализа влияет плотность расположения точек. Для получения сопоставимых результатов кри-

Таблица 2. Таблица расстояний между парами точек $kr = 8.196$

Table 2. Table of distances between pairs of points $kr = 8.196$

i	j																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1		3.22	2.95	4.20	28.05	20.53	11.21	13.69	8.09	6.00	9.53	15.10	9.66	9.59	14.24	12.00	12.74	7.24	17.48
2	3.22		1.69	3.23	30.29	22.75	11.26	14.53	9.98	7.72	10.56	15.84	12.08	11.14	14.91	13.51	10.93	6.46	17.93
3	2.95	1.69		2.39	30.36	21.64	10.09	13.33	9.77	6.91	9.26	14.59	11.94	10.72	13.92	13.11	11.24	5.77	18.50
4	4.20	3.23	2.39		31.53	21.81	8.96	12.69	10.83	7.36	8.58	13.94	12.84	11.38	13.61	13.63	11.37	4.70	19.17
5	28.05	30.29	30.36	31.53		28.18	35.15	33.56	22.37	26.19	31.64	34.63	19.34	27.21	33.83	26.61	36.83	33.64	29.59
6	20.53	22.75	21.64	21.81	28.18		17.50	13.03	20.69	18.71	15.36	13.12	19.05	16.97	13.17	15.71	28.04	21.74	30.39
7	11.21	11.26	10.09	8.96	35.15	17.50		5.20	15.94	11.10	4.28	6.01	17.18	13.71	8.38	14.75	16.00	8.77	24.89
8	13.69	14.53	13.33	12.69	33.56	13.03	5.20		16.77	12.37	5.08	1.66	17.23	14.05	6.71	14.07	19.73	12.77	27.09
9	8.09	9.98	9.77	10.83	22.37	20.69	15.94	16.77		5.13	12.62	18.04	4.13	12.47	18.62	13.74	19.58	14.41	18.62
10	6.00	7.72	6.91	7.36	26.19	18.71	11.10	12.37	5.13		7.90	13.60	7.40	11.81	15.03	13.24	16.96	10.88	20.24
11	9.53	10.56	9.26	8.58	31.64	15.36	4.28	5.08	12.62	7.90		6.31	13.74	12.19	9.34	13.07	17.44	9.89	23.93
12	15.10	15.84	14.59	13.94	34.63	13.12	6.01	1.66	18.04	13.60	6.31		18.58	15.31	7.20	15.24	20.61	13.85	28.37
13	9.66	12.08	11.94	12.84	19.34	19.05	17.18	17.23	4.13	7.40	13.74	18.58		11.99	18.56	12.76	21.11	15.76	18.36
14	9.59	11.14	10.72	11.38	27.21	16.97	13.71	14.05	12.47	11.81	12.19	15.31	10.99		11.20	3.70	18.77	11.58	16.12
15	14.24	14.91	13.92	13.61	33.83	13.17	8.38	6.71	18.62	15.03	9.34	7.20	18.56	11.20		11.05	18.88	12.15	25.10
16	12.00	13.51	13.11	13.63	26.61	15.71	14.75	14.07	13.74	13.24	13.07	15.24	12.76	3.70	11.05		21.23	14.15	18.69
17	12.74	10.93	11.24	11.37	36.83	28.04	16.00	19.73	19.58	16.96	17.44	20.61	21.11	18.77	18.88	21.23		9.62	23.22
18	7.24	6.46	5.77	4.70	33.64	21.74	8.77	12.77	14.41	10.88	9.98	13.85	15.76	11.58	12.15	14.15	9.62		19.04
19	17.48	17.93	18.50	19.17	29.59	30.39	24.89	27.09	18.62	20.24	23.93	28.37	18.36	16.12	25.10	18.69	23.22	19.04	

Таблица 3. Строковые кластеры к примеру Н. С. Ковальчук (9 компонентов, 19 проб, критическое расстояние 8.196, большие расстояния опущены, 28 связей)

Table 3. String clusters, for example, N. S. Kovalchuk (9 components, 19 samples, critical distance 8.196, long distances omitted, 28 links)

i	меры сходства точки i с точкой j																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1		3.22	2.95	4.20					8.09	6.00								7.24	
2	3.22		1.69	3.23						7.72								6.46	
3	2.95	1.69		2.39						6.91								5.77	
4	4.20	3.23	2.39							7.36								4.70	
5																			
6																			
7								5.20			4.28	6.01							
8							5.20				5.08	1.66			6.71				
9	8.09									5.13			4.13						
10	6.00	7.72	6.91	7.36					5.13		7.90		7.40						
11							4.28	5.08		7.90		6.31							
12							6.01	1.66			6.31				7.20				
13									4.13	7.40									
14																3.70*			
15								6.71				7.20							
16														3.70					
17																			
18	7.24	6.46	5.77	4.70															
19																			

* кластер № 2 (точки 14—16) / cluster 2 (14—16)

тические расстояния целесообразно нормировать, как указано в алгоритме.

Как бы удачно ни было выбрано критическое расстояние, оно характеризует сгруппированность точек только для единственного $r_{кр}$. Практика ком-
пьютерного моделирования кластеров в некотором множестве точек [11, 12, 17] и последующего реше-

ния обратной задачи, т. е. проведения кластерного анализа, показали, что нет способа выбора $r_{кр}$, при котором его результаты были бы максимально при-
ближены к замоделированной картине кластеров. Поэтому в разработанном алгоритме предусмотре-
на обработка данных с последовательностью крити-
ческих расстояний, которая позволяла бы постро-

ить кластограмму динамики развития (объединения) кластеров с увеличением $r_{кр}$, подобно тому, как это делается в корреляционном анализе, где изображается объединение пар признаков с увеличением/уменьшением корреляции между ними.

Возникает вопрос о шаге изменения (увеличения) $r_{кр}$, который позволял бы получить наиболее информативную кластограмму. Если последовательность критических расстояний составляет арифметическую прогрессию, т. е. последовательность с равным шагом Δr , то приращиваемое число точек в кластерах будет увеличиваться приблизительно пропорционально nD -й степени критического расстояния, где nD — размерность пространства. Отсюда возникает задача составить шкалу критических расстояний с шагом, уменьшающимся к корню k -й степени из критического расстояния. В этом случае приращение числа точек, объединяемых в кластеры на каждом шаге $r_{кр}$, будет в первом приближении одинаковым. Как рассчитать такую шкалу, любезно показал мне доктор г.-м. наук В. И. Ракин. Это обстоятельство, в частности, позволяет использовать её (шкалу) для проверки статистической гипотезы о том, что *кластеры носят случайный характер*. Вызывает недоумение, что в так называемом стандарте ЮК, изложенном в солидном труде «Основы литохимии», проблема критических расстояний не только не обсуждается, но даже не упоминается [18].

Представление результатов кластерного анализа

Первичной формой представления результатов является таблица, содержащая колонку критических расстояний и колонку одного или нескольких множеств точек, образующих кластеры по числу этих множеств (табл. 4). По этой таблице строится дендрограмма (кластограмма), которая даёт представление о точках, объединяемых в кластеры в зависимости от критических расстояний в динамике (рис. 1). Наконец, строится плоская (двухмерная) проекция расположения точек по их двухмерным координатам, на которой намечаются овальные границы кластеров (рис. 2).

Ни одна из этих форм представления, ни их совокупность не дают исчерпывающего, а также наглядного представления о результатах кластерного анализа. Так, в табл. 4 представлены множества точек, из которых состоят кластеры, т. е. число точек и их номера. Восстановить, какие из этих точек находятся на расстояниях, меньших критического, не представляется возможным. Это невозможно сделать и с помощью дендрограммы (рис. 1). Лишь очень приблизительно это можно сделать по рис. 2, т. е. только для двухмерного случая. Для многомерного случая изобразить пространственное расположение точек можно только

с помощью $C_k^2 = \frac{k!}{2!(k-2)!}$ пар проекций, что для силикатного анализа составляет $\frac{10!}{2 \cdot 8!} = \frac{9 \cdot 10}{2} = 45$ пар. Наглядно представить и осмыслить такое число пар проекций невозможно. К тому же нанесение на проекции границ кластеров — ошибочная процедура, так как *кластер не область пространства, а множество точек*.

Таблица 4. Ранжированные меры сходства между точками в паре (евклидовы расстояния) к примеру Н. С. Ковальчук (9 компонентов, 19 проб, 28 связей, два кластера, критическое расстояние 8.196)

Table 4. Ranked measures of similarity between points in a pair (euclidean distances), for example, N. S. Kovalchuk (9 components, 19 samples, 28 links, two clusters, critical distance 8.196)

Номера точек в паре Number of points in pair	Мера сходства («расстояния») между парами точек Measure of similarity ("distance") between pair of points	Толщина линий, их вид на графе, число связей Width of lines, their appearance on graph, No. of bonds
8-12	1.66	сплошные жирные / bold
2-3	1.69	
3-4	2.39	
1-3	2.95	
1-2	3.22	
2-4	3.23	
14-16	3.70*	
9-13	4.13	
1-4	4.20	
		9 связей / 9 bonds
7-11	4.28	сплошные полужирные / semibold
4-18	4.70	
8-11	5.08	
9-10	5.13	
7-8	5.20	
3-18	5.77	
1-10	6.00	
7-12	6.01	
11-12	6.31	
		9 связей / 9 bonds
2-18	6.46	штриховые / dashed
8-15	6.71	
3-10	6.91	
12-15	7.20	
1-18	7.24	
4-10	7.36	
10-13	7.40	
2-10	7.72	
10-11	7.90	
1-9	8.09	
		10 связей / 10 bonds

* полужирным напечатан кластер № 2

* semibold for cluster 2.

Рис. 1. Изменение числа и состава точек в кластерах в зависимости от критического расстояния (в долях среднего «ближайшего» расстояния между точками), по которому выделялись кластеры (показаны горизонтальными отрезками). А, В, С, D — кусты кластеров

Fig. 1. Changes in number and composition of points in clusters depending on the critical distance (in fractions of average «closest» distance between points) along which clusters were allocated (shown by horizontal lines). A, B, C, D — bushes of clusters

Рис. 2. Дендрограмма кластеров. По оси абсцисс — номера точек, по оси ординат — критические расстояния, по которым выделялись кластеры. А, В, С, D — кусты кластеров

Fig. 2. Dendrogram of clusters. The abscissa axis is number of points, and the ordinate is critical distance along which the clusters were separated. A, B, C, D — bushes of clusters

Таким образом, по обычно представляемым результатам кластерного анализа невозможно определить, является данный кластер полносвязным или неполносвязным, что ограничивает предметный анализ результатов.

Графы, их роль в кластерном анализе

Наиболее близкой к кластерному анализу по целям и методам отраслью науки является *теория графов*. Обе они используют для решения своих задач взаимное расположение точек. В кластерном анализе критерием выделения подмножества точек в кластер служит их взаимная близость, конкретнее — расстояние в многомерном пространстве. Однако «облако» точек в кластере наглядно изобразить и представить невозможно, кроме как в двухмерном пространстве. Вместе с тем невозможно в явном виде установить, какие точки связаны друг с другом близкими расстояниями, меньшими критического, а какие — нет. Такие точки, однако, в неполносвязанном кластере имеют-

ся. Провести полноценный предметный анализ в этих условиях невозможно.

В теории графов понятие *расстояние* отсутствует. Никакая метрика не используется. Граф представляет собой множество точек (вершин) и *рёбер*, которые соединяют пару точек, если между точками этой пары имеется *связь*. Природа этой связи в теории графов не обсуждается, и её наличие (или отсутствие) между указанной парой точек устанавливается из внешних к теории источников. Таким внешним источником для установления наличия (или отсутствия) *связи* может служить кластерный анализ, а именно, если между указанной парой точек расстояние меньше заданного исследователем критического, постулируется связь между ними, иначе постулируется её отсутствие. Теория графов содержит много результатов.

Отсюда непосредственно вытекает, что граф есть наглядное и исчерпывающее представление результатов кластерного анализа. Другими словами, кластерный анализ является источником информации о свя-

Рис. 3. Представление кластеров в форме графов:

1—5 — некоторые типы графов-кластеров: 1 — неполносвязанный простой цикл; 2 — полностью связанный, сумма циклов; 3 — неполносвязанный, радиальный; 4 — неполносвязанный сложный с перешейком; 5 — неполносвязанный, простая цепь. Сплошные линии — связи (рёбра), пунктирная — отсутствующие связи, жирная линия — перешеек

Fig. 3. Representation of clusters in the form of graphs.

1—5 — some types of graph-clusters: 1 — incompletely connected, simple cycle, 2 — completely connected, sum of cycles, 3 — incompletely connected, radial, 4 — incompletely connected, complex with the bridge, 5 — incompletely connected, simple chain. Solid lines — links (edges), dotted — missing links, thick line — bridge

зях между точками, используемыми в теории графов. В свою очередь, теория графов может считаться прямым продолжением кластерного анализа. *Все результаты теории графов остаются верными и применимыми в кластерном анализе.*

В первую очередь легко решается вопрос о наглядном представлении кластера. Им является *конфигурация* графа, совокупность точек (вершин) и связей между ними (*рёбер*), т. е. *сам граф*. Непосредственно по графу видно, является кластер полностью связанным или неполносвязанным. Подсчетом числа вершин и рёбер графа можно выразить *степень связности кластера* $k_{св}$:

$$k_{св} = \frac{l_{фак}}{l_{макс}} = \frac{l_{фак}}{C_n^2}.$$

где $l_{фак}$ — фактическое число рёбер, $l_{макс}$ — максимальное число рёбер при заданном числе точек n .

Теория графов с её многовековой историей имеет многочисленные количественные характеристики графов, такие как длина *маршрута* (*пути*) от одной

вершины до какой-либо другой (выраженная в *числе рёбер*, а не в сумме их длин, так как в теории графов понятия длины ребра не существует!), *диаметр* графа, его *центр* (вершина, сумма путей от которой до остальных вершин минимальна), *контур* графа. Граф считается заданным, если известна его *конфигурация*. Одним из основополагающих концепций теории графов является отсутствие понятия расстояния между вершинами и, соответственно, длины рёбер, т. е. метрики пространства. *Отсутствует само понятие пространства*, координат в нем, его размерности. Тем не менее конфигурация графа может быть задана однозначно *матрицей инцидентности*, или *матрицей смежности*, связывающей в систему сцепление рёбер и вершин. Все это позволяет использовать теорию графов для количественной характеристики формы кластеров, которая в настоящее время находится практически на нуле. С помощью теории графов из кластерного анализа может извлекаться дополнительная информация. В теории графов соединение вершин рёбрами производится, когда вершины графа *взаимодействуют* между собой. Если взаимодействие симметричное, образуется граф *неориентированный*; если оно одностороннее, то образующийся граф называется *ориентированным*.

Ориентированный граф обозначается на рисунках таким образом, что его рёбра снабжаются стрелками, направленными от действующей вершины к вершине, которая испытывает воздействие. Источник информации о направлении взаимодействия, если оно имеется, в теории графов черпается извне, как и информация о наличии *связи*.

Поначалу казалось, что в применении к кластерному анализу ориентированных графов быть не может. Однако это не так. Каждой точке может быть приписано одно из D по числу признаков или некоторая функция из некоторого числа признаков по замыслу исследователя. Например, в геохимии, по данным силикатного анализа, каждой точке может быть приписана сумма щелочей. Тогда стрелки ориентированного графа будут везде направлены от точки с большей суммой к точке с меньшей, как бы указывая пути миграции этих элементов.

В заключение отметим, что в теории графов разработаны многочисленные признаки графов, позволяющие подробно их классифицировать. На рис. 3 приведены некоторые таксоны графов, для которых читатель без труда подберет предметную интерпретацию из своей отрасли.

Литература

1. Богусевский Э. М., Канцель А. В., Коновалов И. М., Певзнер М. Н., Рожкова Г. Ф., Стороженко С. А. Проблема использования вероятностно-статистических характеристик руд для классификации и прогнозирования рудных месторождений // Применение математических методов в геологии. Алма-Ата: Наука, 1969. С. 140—141.
2. Бондаренко В. Н. Принципы и методы построения автоматизированной процедуры идентификации ассоциаций элементов-индикаторов // Автоматиз. анал. и обраб. данных геохим. поисков месторожд. М., 1987. С. 5—18.
3. Бугаец А. Н., Дорофеев А. А., Мацак А. П. К вопросу о некоторых формальных методах и задаче классификации

ции и разграничения геологических объектов // Геол. формации. Л., 1968.

4. Валиахметова А. Д. Кластерный анализ и классификация в геологии // Севергеоэкотех-2000: Межрегиональная молодежная научная конференция. Ухта: Ухтинский государственный технический университет, 15–17 марта 2000.

5. Воронин Ю. А. Основные результаты по простой теории классифицирования в геологии // Вычисл. центр СО АН СССР. Препр. 1981. № 315. 29 с.

6. Воронин Ю. А. Начала теории сходства. Новосибирск: ВЦ СО АН СССР, 1989. 120 с.

7. Воронин Ю. А., Гольдина Н. А. Пример совместного упрощенного математико-логического разбора геологических классификаций // Геология и геофизика. 1965. № 2. 4 с.

8. Горбатов В. А. Основы дискретной математики: Учебное пособие для студентов вузов. М.: Высшая школа, 1986. 311 с.

9. Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П. Об одном принципе классификации и прогнозирования геологических объектов и явлений // Геология и геофизика. 1968. № 5. С. 50–64.

10. Иванов Б. Н. Дискретная математика. Алгоритмы и программы: Учеб. пособие. М.: Лаборатория базовых знаний, 2003. 288 с.

11. Кириллова В. В., Ткачев Ю. А. Кластер: от термина свободного пользования к статистически обоснованному объекту // Современные проблемы теоретической, экспериментальной и прикладной минералогии (Юшкинские чтения — 2013): Материалы минералогического семинара с международным участием. Сыктывкар: Геопринт, 2013. С. 385–388.

12. Кириллова В. В., Ткачев Ю. А. Статистика кластеров в совокупности случайно размещенных точек // Там же. С. 388–390.

13. Корнеева А. В., Карлинская Р. Р. Кластерный анализ бинарных признаков и признаков, характеризующих качествами грациями // Тр. Тбил. ун-та. 1987. 272. С. 229–243.

14. Кузнецов О. П., Адельсон-Вельский Г. М. Дискретная математика для инженера. М.: Энергоатомиздат, 1988. 480 с.

15. Новые идеи в научной классификации: коллективная монография. Вып. 5. Екатеринбург: УрО РАН, 2010. 631 с.

16. Покровский М. П. Введение в классиологию. Екатеринбург: ИГГ УрО РАН, 2014. 484 с.

17. Ткачев Ю. А. Структурно-статистический анализ пространственного взаимного расположения геологических объектов // Уральский геологический журнал. 2010. № 5. С. 53–62.

18. Юдович Я. Э., Кетпис М. П. Основы литохимии. СПб.: Наука, 2000. 479 с.

19. Botbol Joseph Moses. Multivariate clustering based on entropy / US Geol. Surv. Bull. 1989. № 1893, 1–30.

20. Hawrins Douglas M., Rasmussen Stephen E. <Использование дискриминантного анализа для классификации стратиграфических горизонтов в осадочных породах> J. Int. Assoc. Math. Geol. 1973, 5, №2, 163–177.

21. Men Gui Zhen. The statistical analysis in the classification of bituminous coal. «Horn: pribram vede a techn.: Symp. prac. ban. prum.», 1987. Sek.: Mat. metody geol. Sb. Dil MI». S. 1., 1987, 182–188.

22. Woronow Alex. Cluster analysis. «Geobyte», 1987, 2, № 4, 34–37.

References

1. Bogushevskii E. M., Kantsel' A. V., Kononov I. M., Pevzner M. N., Rozhkova G. F., Storozhenko S. A. *Problema ispolzovaniya veroyatnostno-statisticheskikh harakteristik rud dlya klassifikatsii i prognozirovaniya rudnykh mestorozhdenii* (The problem of using probabilistic-statistical characteristics of ores for the classification and forecasting of ore deposits. In: "Application of mathematical methods in geology). Alma-Ata: Nauka, 1969, pp. 140–141.

2. Bondarenko V. N. *Printsipy i metody postroeniya avtomatizirovannoi protsedury identifikatsii assotsiatsii elementov-indikatorov*. "Avtomatiz. anal. i obrab. dannykh geohim. poiskov mestorozhd" (Principles and methods for constructing an automated procedure for identifying associations of indicator elements. "Automation. anal and processing geochemical data prospecting field). Moscow, 1987, pp. 5–18.

3. Bugaets A. N., Dorofeyuk A. A., Matsak A. P. *K voprosu o nekotorykh formal'nykh metodakh i zadache klassifikatsii i razgranicheniya geologicheskikh ob'ektov* (Some formal methods and the task of classifying and delimiting geological objects). In: Geol. formatsii. Leningrad, 1968, pp. 00–00.

4. Valiahetova A. D. *Klasternyi analiz i klassifikatsiya v geologii*. (Uhtinskii gosudarstvennyi tekhnicheskii universitet). *Mezhregional'naya molodezhnaya nauchnaya konferentsiya "Severgeokotech-2000"* (Cluster analysis and classification in geology. (Ukhta State Technical University). Interregional Youth Scientific Conference "Severgeokotech-2000"), Ukhta, 15–17 March, 2000.

5. Voronin Yu. A. *Osnovnye rezul'taty po prostoi teorii klassifitsirovaniya v geologii* (Main results on a simple classification theory in geology). "Vychisl. tsentr SO AN SSSR. Prepr.", 1981, No. 315, 29 pp.

6. Voronin Yu. A. *Nachala teorii shodstva* (Principles of the theory of similarity). Novosibirsk, SB AS USSR, 1989, 120 pp.

7. Voronin Yu. A., Goldina N. A. *Primer sovместnogo uproschenno matematiko-logicheskogo razbora geologicheskikh klassifikatsii* (An example of a joint simplified mathematical-logical analysis of geological classifications). *Geologiya i geofizika*, 1965, No. 2, 4 pp.

8. Gorbato V. A. *Osnovy diskretnoi matematiki: Uchebnoe posobie dlya studentov vuzov* (Basics of Discrete Mathematics: A manual for university students). Moscow: Vysshaya shkola, 1986, 311 pp.

9. Dmitriev A. N., Zhuravlev Yu. I., Krendel'ev F. P. *Ob odnom printsipe klassifikatsii i prognozirovaniya geologicheskikh ob'ektov i yavlenii* (On a principle of classification and forecasting of geological objects and phenomena). *Geologiya i geofizika*, 1968, No. 5, pp. 50–64.

10. Ivanov B. N. *Diskretnaya matematika. Algoritmy i programmy* (Discrete Mathematics. Algorithms and programs). Moscow: *Laboratoriya bazovykh znaniy*, 2003, 288 pp.

11. Kirillova V. V., Tkachev Yu. A. *Klaster: ot termina slobodnogo pol'zovaniya k statisticheski obosnovannomu ob'ektu/ Sovremennye problemy teoreticheskoi, eksperimental'noi i prikladnoi mineralogii* (Yushkinskie chteniya — 2013) (Cluster: from the term of free use to a statistically valid object / Modern problems of theoretical, experimental and applied mineralogy (Yushkin readings — 2013). Proceedings of the mineralogical seminar with international participation. Syktvykar: Geoprint, 2013, pp. 385–388.

12. Kirillova V. V., Tkachev Yu. A. *Statistika klasterov v sovokupnosti sluchaino razmeshchennykh tocheek/ Sovremennye problemy teoreticheskoi, eksperimental'noi i prikladnoi min-*

eralogii (*Yushkinskie chteniya* — 2013). (Statistics of clusters in aggregate of randomly placed points. Modern problems of theoretical, experimental and applied mineralogy (Yushkin readings — 2013). Proceedings of the mineralogical seminar with international participation. Syktyvkar: Geoprint, 2013, pp. 388—390.

13. Korneeva A. V., Karlinskaya R. R. *Klasternyi analiz binarnykh priznakov i priznakov, harakterizuyemykh kachestvennyimi gradatsiyami* (Cluster analysis of binary signs and signs characterized by qualitative gradations). Tr. Tbil. un-ta, 1987, 272, pp. 229—243.

14. Kuznetsov O. P., Adelson-Velsky G. M. *Diskretnaya matematika dlya inzhenera* (Discrete mathematics for an engineer). Moscow: Energoatomizdat, 1988, 480 pp.

15. *Novye idei v nauchnoi klassifikatsii: kollektivnaya monografiya* (New ideas in scientific classification: a collective monograph). Issue 5. Ekaterinburg: UB RAS, 2010, 631 pp.

16. Pokrovskii M. P. *Vvedenie v klassiologiyu* (Introduction to the classification). Ekaterinburg: IGG UB RAS, 2014, 484 pp.

17. Tkachev Yu. A. *Strukturno-statisticheskii analiz prostanstvennogo vzaimnogo raspolozheniya geologicheskikh ob'ektov* (Structural and statistical analysis of the spatial mutual arrangement of geological objects). Uralsky geologicheskii zhurnal, No. 5, 2010, pp. 53—62.

18. Yudovich Ya. E., Ketris M. P. *Osnovy litohimii* (Basics of lithochemistry). St. Petersburg: Nauka, 2000, 479 pp.

19. Botbol Joseph Moses. Multivariate clustering based on entropy. US Geol. Surv. Bull. 1989. No 1893, 1—30.

20. Hawrins Douglas M., Rasmussen Stephen E. <Ispol'zovanie diskriminantnogo analiza dlya klassifikatsii stratigraficheskikh gorizontov v osadochnykh porodakh> J. Int. Assoc. Math. Geol. 1973, 5, No.2, 163—177.

21. Men Gui Zhen. The statistical analysis in the classification of bituminous coal. “Horn: pribram vede a techn.: Symp. prac. ban. prum., 1987. Sek.: Mat. metody geol. Sb. Dil MI”. S. 1., 1987, 182—188.

22. Woronow Alex. Cluster analysis. “Geobyte”, 1987, 2, No. 4, pp. 34—37.