КЛАСТЕРНЫЙ АНАЛИЗ В ГЕОЛОГИИ: МЕТРИКИ И МЕРЫ

Ю. Л. Войтеховский

Санкт-Петербургский горный университет, Санкт-Петербург

Voytekhovskiy_YuL@pers.spmi.ru

Статья посвящена кластерному анализу, активно применяемому при статистической обработке данных в геологии, биологии и других естественных науках. Обращено внимание на то, что результат кластерного анализа может существенно зависеть от применяемой метрики, т.е. расстояния между точками в многомерном пространстве. Расстояния между точечными множествами могут определяться непосредственно и через меры покрывающих множеств. Методический ресурс кластерного анализа видится в сравнении результатов при использовании разных мер и метрик.

Ключевые слова: кластерный анализ, метрики, меры.

CLUSTER ANALYSIS IN GEOLOGY: METRICS AND MEASURES

Yu. L. Voytekhovsky

Saint-Petersburg Mining University, Saint-Petersburg

Voytekhovskiy_YuL@pers.spmi.ru

The article is devoted to cluster analysis, actively used in statistical data processing in geology, biology, and other natural sciences. Attention is drawn to the fact that the result of cluster analysis may significantly depend on the metric used, i.e. distances between points in multidimensional space. The distances between the point sets can be determined directly and by the measures of covering sets. The methodical resource of cluster analysis is seen in comparing the results when using different measures and metrics.

Keywords: cluster analysis, metrics, measures.

Введение

Недавняя статья [2] обратила внимание читателей на неисчерпанный методический ресурс кластерного анализа, активно применяемого в естественных науках. Справедливо замечено, что именно введение метрики (расстояния) превратило его в количественный анализ распределения точек в многомерном пространстве. В то же время следует подчеркнуть, что метод диалектичен: выделяя кластеры точек (что бы ни понималось под таковыми), он не только классифицирует (разделяет), но и связывает подмножества исходного точечного множества в иерархически упорядоченную структуру с нестрогим отношением порядка. В этой статье обращено внимание на то, что результат кластерного анализа существенно зависит от применяемой метрики, а расстояния между кластерами можно определять непосредственно или через меры покрывающих множеств.

Метрики

Множество называется метрическим пространством, если для любых двух его элементов i и j определено вещественное число h_{ij} , называемое метрикой (расстоянием) и обладающее свойствами: (1, неотрицательность) $h_{ij} \geq 0$, причём $h_{ij} = 0$ тогда и только тогда, когда $i = j$; (2, симметричность) $h_{ij} = h_{ji}$; (3, неравенство треугольника) для любых трёх элементов — i , j и k — выполнено $h_{ij} + h_{jk} \geq h_{ik}$ [1].

На одном множестве могут быть заданы разные метрики, порождающие разные метрические пространства. Так, расстояния между точками на плоскости можно измерить привычной евклидовой (впрочем, столь же пифагоровой и декартовой, поскольку

основана на теореме Пифагора в декартовой ортогональной системе координат) метрикой:

$$h_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2}.$$

Непривычна, но корректна и другая метрика:

$$h_{ij}^* = |x_i - x_j| + |y_i - y_j|.$$

Здесь расстояние измеряется не длиной гипотенузы, а суммой длин катетов того же треугольника. Выполнение свойств 1 и 2 очевидно. Свойство 3 легко проверяется для различных взаимных положений точек. Обе метрики обобщаются на многомерные пространства.

Ситуация получает новый акцент, если нужно найти расстояние между двумя точечными множествами, в том числе перекрывающимися по координатам точек. Нередко считается, что оно равно расстоянию между ближайшими точками множеств. Ошибочность такого решения видна из примера. Пусть совпадают две точки, принадлежащие различным множествам. Тогда расстояние между ними равно нулю. Но свойство 1 требует, чтобы в этом случае множества совпадали тождественно, т.е. всеми своими точками. Легко показать, что при указанном (неправильном) способе измерения нарушается и свойство 3. Между тем он практикуется, например, во всех компьютеризованных анализаторах петрографических структур, где под расстоянием между минеральными зёрнами понимается евклидово расстояние между точками (например, геометрическими центрами тяжести), взятыми в пределах сечений.

Для измерения расстояний между точечными множествами годится на первый взгляд странная, устроенная иерархически, рассчитываемая поэтапно минимаксная метрика Хаусдорфа:

$$H(I, J) = \max \{ \{ \max \zeta(i, J) \mid i \in I \}, \{ \max \zeta(j, I) \mid j \in J \} \},$$

где $\zeta(i, J) = \min \{ h_{ij} \mid j \in J \}$, $\zeta(j, I) = \min \{ h_{ji} \mid i \in I \}$ (рис. 1). Здесь $\max \zeta(i, J) \mid i \in I$ — длина отрезка $[1, 1']$, $\max \zeta(j, I) \mid j \in J$ — длина отрезка $[2, 2']$. Расстояние между точечными множествами I и J равно длине отрезка $[2, 2']$ как большей из двух. Вместо h_{ij} можно использовать h_{ij}^* или другую метрику, заданную на парах точек.

Рис. 1. Метрика Хаусдорфа
Fig. 1. The Hausdorff metric

Меры

Множество элементов называется измеримым пространством, если на его подмножествах (вообще говоря, образующих алгебру, но мы не будем углубляться в эти тонкости) определена вещественная ограниченная функция, называемая мерой и обладающая свойствами: (1, неотрицательность): $\mu(I) \geq 0$; (2, монотонность): если $I \subset J$, то $\mu(I) \leq \mu(J)$; (3, аддитивность): если $I \cap J = \emptyset$, то $\mu(I \cup J) = \mu(I) + \mu(J)$ [1].

Определение метрики через меры возможно и целесообразно, когда точечное множество тем или иным способом оконтуривается, т. е., по сути, покрывается фигурой с ненулевой мерой. В качестве метрик можно предложить:

$$\rho(I, J) = \mu(I \cup J) - \mu(I \cap J)$$

$$\text{и } \rho(I, J)^* = 1 - \mu(I \cap J) / \mu(I \cup J).$$

В обоих случаях относительную трудность представляет проверка свойства 3 метрики. На рис. 2 для простоты точечные множества I и J оконтурены прямоугольниками. Здесь под мерами естественно понимать площади их объединения и пересечения, вычисляемые по координатам крайних точек:

$$\begin{aligned} \mu(I \cup J) = & |\max x_i - \min x_i| \times |\max y_i - \min y_i| + \\ & + |\max x_j - \min x_j| \times |\max y_j - \min y_j| - \\ & - |\max x_j - \min x_i| \times |\max y_j - \min y_i|, \end{aligned}$$

$$\mu(I \cap J) = |\max x_j - \min x_i| \times |\max y_j - \min y_i|.$$

Рис. 2. Определение метрики через меры
Fig. 2. Definition of metric by measure

Заключение

Сказанное может породить вопросы: какую метрику следует применить в том или ином случае, зависит ли выбор главным образом от предметного содержания задачи или от конфигурации точечного множества. По-видимому, категорических ответов на них нет. С одной стороны, математика бесстрашна, как скальпель хирурга. С другой стороны, применяя математические методы, хорошо бы как можно лучше знать специфику изучаемого явления. Лучшим решением представляется последовательное применение в кластерном анализе нескольких метрик. Они редко приводят к полному совпадению результатов. Здесь и наступает самое интересное — попытка содержательной интерпретации различий, через которые просвечивает нечто искомое.

Литература

1. Вулих Б. З. Краткий курс теории функций вещественной переменной. М.: Наука, 1965. 304 с.
2. Ткачев Ю. А. Основания многомерного кластерного анализа в геологии // Вестник Института геологии Коми НЦ УрО РАН. 2019. № 2. С. 44—52. doi: 10.19110/2221-1381-2019-2-44-52.

References

1. Vulikh B. Z. *Kratkiy kurs teorii funktsiy veshchestvennoy peremennoy* (Short course of the theory of functions of a real variable). B. Z. Vulikh. Moscow: Nauka, 1965, 304 pp.
2. Tkachev Yu. A. *Osnovaniya mnogomernogo klaster'nogo analiza v geologii* (Foundations of multidimensional cluster analysis in geology). Yu. A. Tkachev. Vestnik of institute of geology Komi SC UB RAS, 2019, No. 2, pp. 44—52. doi: 10.19110/2221-1381-2019-2-44-52.