УДК 519.25

С.В. Дронов

Одна кластерная метрика и устойчивость кластерных алгоритмов

S.V. Dronov

A Cluster Metric and Stability of Cluster Algorithms

Одной из проблем, которые возникают при решении задачи кластерного анализа, является сравнение нескольких решений этой задачи, получающихся при использовании различных методов или меняющихся стартовых конфигураций. Предложен подход к сравнению двух кластерных разбиений одного и того же конечного множества. На основе вводимой для этого метрики на классе всех возможных разбиений множества на кластеры определяется и обсуждается понятие устойчивости кластерных алгоритмов.

Ключевые слова: кластерное разбиение, устойчивость, метрика на конечном множестве.

One of the problems arising in cluster analysis is a problem of comparing several different cluster partitions. The cause of such problems may lie in using more than one method of clusters building or trying the same method with different starting configurations. In the paper we suggest an approach to comparing two cluster partitions of the same finite set. It is based on some metric on the class of all cluster partitions. A notion of stability for cluster algorithms is also defined and discussed.

Key words: cluster partition, stability, metric on finite set.

1. Кластерная метрика. Пусть X – конечное множество. Для произвольного $A \subset X$ через $|A|$ будем обозначать количество его элементов. Класс $\mathcal{A}$ подмножеств X назовем кластерным разбиением, если каждое $x \in X$ является элементом некоторого множества $A_x \in \mathcal{A}$ и все множества, из которых состоит $\mathcal{A}$, попарно дизъюнктны, т.е.

$$(\forall A, B \in \mathcal{A})\ (A \neq B) \Rightarrow (A \cap B = \emptyset).$$

Для двух кластерных разбиений $\mathcal{A}, \mathcal{B}$ одного и того же множества X из n элементов определим

$$d(\mathcal{A}, \mathcal{B}) = \sum_{x \in X} |A_x \Delta B_x|,$$

где $A \Delta B = (A \cup B) \setminus (A \cap B)$ – симметрическая разность множеств. Нетрудно доказать, что введенная величина представляет собой метрику на наборе всевозможных кластерных разбиений X.

Пусть $\mathcal{A} = \{A_1, ...A_k\}$, $\mathcal{B} = \{B_1, ..., B_m\}$. Введем обозначения

$$\cap_{i,j} = |A_i \cap B_j|, \quad \cup_{i,j} = |A_i \cup B_j|,$$

$$i = 1, ..., k, \quad j = 1, ..., m.$$

Теорема 1.

$$d(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{k} \sum_{j=1}^{m} \cap_{i,j} (\cup_{i,j} - \cap_{i,j}) \ . \qquad (1)$$

Доказательство. Для каждого $x \in A_i \cap B_j$ справедливо $|A_x \Delta B_x| = \cup_{i,j} - \cap_{i,j}$, поэтому достаточно в определении метрики объединить одинаковые по величине слагаемые.

Полезным для контроля вычислений по формуле (1) может оказаться тот факт, что множества $A_i \cap B_j$, $i = 1, ...k$, $j = 1, ..., m$ также образуют кластерное разбиение, откуда получаем

$$\sum_{i=1}^{k} \sum_{j=1}^{m} \cap_{i,j} = n,$$

$$(\forall i, j) \ \sum_{j=1}^{m} \cap_{i,j} = |A_i|, \quad \sum_{i=1}^{k} \cap_{i,j} = |B_j|.$$

Обозначим элементы множества X через $x_1, ..., x_n$, а два наиболее различных (точное содержание этого термина будет ясно чуть ниже) его кластерных разбвиения следующим образом:

$$\underline{X} = \{\{x_1\}, ..., \{x_n\}\}, \quad \overline{X} = \{\{x_1, ..., x_n\}\} = \{X\}.$$

Разбиения $\mathcal{A}, \mathcal{B}$ будем называть совпадающими (равными), если $(\forall A \in \mathcal{A})(\exists B \in \mathcal{B})\ A = B$.

Теорема 2. *Для любых двух кластерных разбиений множества X справедливо неравенство*

$$d(\mathcal{A},\mathcal{B}) \;\le\; n(n-1), \qquad (2)$$

причем равенство достигается тогда и только тогда, когда одно из разбиений совпадает с $\underline{X}$, а другое с $\overline{X}$.

Доказательство. Если $\cap_{i,j} \neq 0$, то, очевидно, $\cup_{i,j} - \cap_{i,j} \le n-1$. При этом равенство здесь достигается тогда и только тогда, когда

$$\cup_{i,j} = n, \quad \cap_{i,j} = 1. \qquad (3)$$

Отметим, что если (3) для каких-то i,j выполнено, то

$$(\exists x \in X)\; A_i \cap B_j = \{x\}, \quad A_i \cup B_j = X.$$

Теперь из (1) получаем

$$d(\mathcal{A},\mathcal{B}) \le (n-1)\sum_{i=1}^{k}\sum_{j=1}^{m}\cap_{i,j} = (n-1)n.$$

Неравенство (2) доказано. Заметим, что если хотя бы для одной пары индексов i,j такой, что $\cap_{i,j} \neq 0$, условие (3) нарушается, то неравенство (2) становится строгим.

Проверка того, что в (2) достигается равенство, если $\mathcal{A} = \underline{X}, \mathcal{B} = \overline{X}$ или наоборот, тривиальна. Обратно, пусть равенство достигнуто, т.е. (3) выполнено для всех i,j таких, что $\cap_{i,j} \neq 0$. Предположим, что в разбиениях $\mathcal{A},\mathcal{B}$ нашлись хотя бы по два множества (A_1, A_2 и B_1, B_2 соответственно – это означает, что ни одно из разбиений не равно $\overline{X}$), причем множества A_1, B_1 содержат более чем по одному элементу. Последнее условие эквивалентно тому, что ни одно из разбиений не совпадает с $\underline{X}$.

Случай 1. Пусть $A_1 \cap B_1 \neq \emptyset$. Тогда это пересечение состоит из единственного элемента x и $A_1 \cup B_1 = X$, откуда $B_2 \subset A_1 \setminus \{x\}$. Поскольку A_1 не одноэлементно, значит $A_1 \cap B_2 \neq \emptyset$, т.е. $A_1 \cup B_2 = X$ согласно (3). Но $B_1 \cap B_2 = \emptyset$, откуда следует, что $B_1 = \{x\}$. Противоречие.

Случай 2. $A_1 \cap B_1 = \emptyset$. Без ограничения общности можно считать, что все элементы $\mathcal{B}$, имеющие непустые пересечения с A_1 (а такие есть), одноэлементны. Пусть $B_2 = \{z\}$ такое множество. Тогда $B_2 \subset A_1$ и с учетом (3) получаем

$$A_1 \cup B_2 = A_1 = X \;\Rightarrow \mathcal{A} = \overline{X},$$

или состоит лишь из одного элемента. Вновь противоречие.

Если $\mathcal{A} = \underline{X}$ и равенство в (2) достигнуто, то в предположении, что $A_i \cap B_j \neq \emptyset$, немедленно получается $A_i \cup B_j = B_j$, то есть из (3) $B_j = X$, откуда $\mathcal{B} = \overline{X}$. Если же $\mathcal{A} = \overline{X}$, то при произвольных i,j справедливо $A_i \cap B_j = B_j$. Вновь используя (3), приходим к выводу, что все B_j одноэлементны, т.е. $\mathcal{B} = \underline{X}$. Теорема доказана.

Основываясь на доказанной теореме, можно ввести коэффициент кластерных различий:

$$K(\mathcal{A},\mathcal{B}) \;=\; \frac{d(\mathcal{A},\mathcal{B})}{n(n-1)}\,.$$

Он принимает значения между 0 и 1, и, чем меньше он по величине, тем более похожими друг на друга являются кластерные разбиения.

2. Другой способ расчета введенной метрики. При попытке вычислять значения введенной метрики по формуле (1) на практике сразу возникают значительные трудности. Предложим другой способ вычисления d.

Лемма 1. *Пусть для произвольных i,j*

$$T_{i,j} \;=\; \sum_{r\neq j}\cap_{i,r} \;+\; \sum_{s\neq i}\cap_{s,j}.$$

Тогда имеет место формула

$$d(\mathcal{A},\mathcal{B}) \;=\; \sum_{i=1}^{k}\sum_{j=1}^{m}\cap_{i,j}T_{i,j}. \qquad (4)$$

Доказательство. Формула (4) немедленно следует из (1), если заметить, что

$$\cup_{i,j} = \sum_{r=1}^{m}\cap_{i,r} + \sum_{s=1}^{k}\cap_{s,j} - \cap_{i,j}$$

для $i = 1,...,k,\; j = 1,...,m$.

Из этой леммы вытекает следующий способ вычисления рассматриваемой метрики. Поместим все числа $\cap_{i,j}$ в матрицу P размерности $k \times m$. Тогда $T_{i,j}$ – сумма элементов этой матрицы, образующих крестообразную фигуру с центром (i,j) без центрального элемента. Вычисляя все такие суммы и помещая их в новую матрицу T, нетрудно завершить вычисления по формуле (4). Этот способ вычисления представляется более алгоритмичным, чем использование формулы (1) и, тем более, определения $d(\mathcal{A},\mathcal{B})$.

Лемма 2. *Пусть разбиение $\mathcal{B}$ получено из разбиения $\mathcal{A}$ перемещением одного элемента*

из множества A_r, $|A_r| = n_r$ *в множество* A_s, $|A_s| = n_s$, $r \neq s$. *Тогда*

$$d(\mathcal{A}, \mathcal{B}) \;=\; 2(n_r + n_s - 1).$$

Доказательство. Понятно, что

$$\cap_{r,r} = n_r - 1,\ \cap_{r,s} = 1,\ \cap_{s,r} = 0,\ \cap_{s,s} = n_s,$$

$$T_{r,r} = 1,\ T_{r,s} = T_{s,r} = n_r + n_s - 1,\ T_{s,s} = 1,$$

а если хотя бы одно из i, j не попадает в множество $\{r, s\}$, то $\cap_{i,j} T_{i,j} = 0$. Утверждение леммы немедленно следует из формулы (4).

3. О расстояниях между k-разбиениями. В некоторых алгоритмах кластерного анализа (например, в методе k-средних) могут изменяться лишь составы кластеров, а число первоначально заданных кластеров не может меняться. Рассмотрим этот случай подробнее. Зафиксируем натуральное число k, меньшее n. Кластерное разбиение $\mathcal{A}$ будем называть k-разбиением, если оно состоит из k множеств.

Пусть $\mathcal{A}, \mathcal{B}$ – два k-разбиения. Для фиксированного $q \in \{1, ..., n\}$ через $\mathcal{B}_q^+$ обозначим набор тех $B_j \in \mathcal{B}$, для которых $\cap_{q,j} \neq 0$.

Лемма 3. *Для произвольного* q *справедливо* $\mathcal{B}_q^+ \neq \emptyset$. *Если при каждом* q *набор* $\mathcal{B}_q^+$ *одноэлементен, то* k*-разбиения* $\mathcal{A}, \mathcal{B}$ *совпадают.*

Доказательство. Первое утверждение – тривиальное следствие определений. Пусть

$$(\forall q)\ (\exists p(q))\ \mathcal{B}_q^+ = \{B_{p(q)}\}.$$

Это означает, что

$$(\forall q = 1, .., k) \quad A_q \subset B_{p(q)}. \tag{5}$$

Если бы двум разным q соответствовало бы одно и то же p, то в силу совпадения количеств множеств в разбиениях какому-то множеству B_s не соответствовало бы ни одного q, т.е. ни одно из $A_q \in \mathcal{A}$ не имело бы с B_s общих элементов, что противоречит определению кластерного разбиения.

Если бы хотя бы при одном q нашелся какой-то элемент $x \in B_{p(q)} \setminus A_q$, то, поскольку при $p \neq p(q)$ справедливо $B_p \cap B_{p(q)} = \emptyset$, то из (5) получалось бы, что

$$\cup_{i=1}^k A_i \subset \cup_{i \neq q} B_{p(i)} \cup A_q \subset X \setminus \{x\},$$

что вновь невозможно. Итак, доказано, что $(\forall q)(\exists p)\ A_q = B_p$, а это означает, что $\mathcal{A} = \mathcal{B}$. Лемма доказана.

Теорема 3. *Пусть* $\mathcal{A}$ *–* k*-разбиение* X. *Тогда ближайшее к нему в смысле* d k*-разбиение, не совпадающее с ним, получается перенесением одного* $x \in X$ *из некоторого элемента* $A \in \mathcal{A}$ *в другой его элемент.*

Доказательство. Пусть $\mathcal{B}_0$ – произвольное k-разбиение X, отличное от $\mathcal{A}$. Выберем такое $A_q \in \mathcal{A}$, что набор $\mathcal{B}_q^+$ хотя бы двухэлементен. Среди всех элементов этого набора B_p найдем такое B_r, что $\cap_{q,r}$ наибольшее. Пусть B_p – другой элемент этого набора и $x \in A_q \cap B_p$. В силу сделанного выбора

$$0 < \cap_{q,p} \leq \cap_{q,r}, \quad \sum_{i \neq q} \cap_{i,r} \leq \sum_{i \neq q} \cap_{i,p}. \tag{6}$$

Перенесем x из B_p в B_r. Новое кластерное разбиение обозначим через $\mathcal{B}_1$. Пусть $d_i = d(\mathcal{A}, \mathcal{B}_i)$, $i = 0, 1$. Все величины, участвующие в формуле (1) при вычислении d_i, будем помечать верхним индексом i.

Заметим, что при переходе от вычисления d_0 к d_1 поменяются указанным ниже образом лишь следующие величины:

$$\cup_{i,r}^1 = \cup_{i,r}^1 + 1, \quad \cup_{i,p}^1 = \cup_{i,p}^0 - 1,\ i \neq q,$$

$$\cap_{q,p}^1 = \cap_{q,p}^0 - 1, \quad \cap_{q,r}^0 + 1.$$

Отсюда немедленно вытекает, что

$$d_1 = d_0 - \textstyle\sum_{i \neq q} \cap_{i,p}^0 + \sum_{i \neq q} \cap_{i,r}^0 - \cup_{q,p}^0 + \cap_{q,p}^0 +$$
$$+ \cup_{q,r}^0 - 2 = d_0 + 2\left(\textstyle\sum_{i \neq q} \cap_{i,r}^0 - \sum_{i \neq q} \cap_{i,p}^0\right) +$$
$$+ \cap_{q,p}^0 - \cap_{q,r}^0 - 2.$$

Из условий (6) получаем

$$d_1 \;\leq\; d_0 - 2 \;<\; d_0,$$

т.е. разбиение $\mathcal{B}_1$ ближе к $\mathcal{A}$, чем $\mathcal{B}_0$. Будем повторять проведенное рассуждение, заменяя в нем разбиение $\mathcal{B}_0$ на $\mathcal{B}_1$, $\mathcal{B}_2, ...$ В силу конечности всех рассматриваемых объектов процесс рано или поздно остановится, причем произойдет это тогда, когда ни для одного q мы не сможем найти двух различных элементов в $\mathcal{B}_q^+$, что, согласно лемме 3, может произойти лишь в случае $\mathcal{B}_j = \mathcal{A}$. Тогда разбиение $\mathcal{B}_{j-1}$ расположено от $\mathcal{A}$ не дальше начального $\mathcal{B}_0$ и отличается от $\mathcal{A}$ лишь положением одного элемента x. В силу произвольности k-разбиения $\mathcal{B}_0$ теорема доказана.

Из теоремы 3 и леммы 2 следует

Теорема 4. *Упорядочим множества, составляющие* k*-разбиение* $\mathcal{A}$, *по возрастанию количеств их элементов и перенумеруем их в этом порядке. Пусть* j *– наименьший номер множества с числом элементов* $n_j \geq 2$. *Если*

$j = 1$, то ближайшее в смысле метрики d к $\mathcal{A}$ k-разбиение, не совпадающее с ним, удалено от него на расстояние

$$d = 2(n_{j+1} + n_j - 1) \ \geq \ 4n_j - 2.$$

Если же $j \geq 2$, то на $d = 2n_j$. В частности, два ближайших различных k-разбиения не могут быть удалены друг от друга менее чем на $d = 4$.

Единственное разбиение, для которого заключение теоремы не работает, – $\underline{X}$. Но этот случай для нас не представляет интереса, поскольку такое разбиение мы не можем трансформировать без изменения количества множеств, его составляющих.

4. Об устойчивости кластерных алгоритмов. В кластерном анализе иногда рассматриваются итерационные процедуры, которые, стартуя с некоторого начального кластерного разбиения $\mathcal{A}$ и как-то трансформируя его, в итоге приходят к окончательному разбиению. Это разбиение, следовательно, может рассматриваться как функция от начального, что можно записать как $\mathcal{B} = F(\mathcal{A})$. Если независимо от начального разбиения результат процедуры всегда получается один и тот же, то естественно назвать такую процедуру абсолютно устойчивой (по отношению к начальному разбиению).

Если же возможны различные результаты в зависимости от начального разбиения, то может оказаться полезным следующее определение. Кластерный алгоритм F описанного типа будем называть ε-устойчивым в точке $\mathcal{A}$, если найдется такое $\delta > 0$, что для любого кластерного разбиения $\mathcal{A}_1$ из условия $d(\mathcal{A}, \mathcal{A}_1) \leq \delta$ следует $d(F(\mathcal{A}), F(\mathcal{A}_1)) \leq \varepsilon$. Можно ввести также определение равномерной ε-устойчивости:

$$(\exists \delta > 0) \ (\forall \mathcal{A}, \mathcal{A}_1) \ (d(\mathcal{A}, \mathcal{A}_1) \leq \delta) \Rightarrow$$
$$\Rightarrow (d(F(\mathcal{A}), F(\mathcal{A}_1)) \leq \varepsilon),$$

а также другие подобные определения.

Нетрудно заметить, что из полученных выше результатов следует, что если при $\varepsilon = 1$ $\delta = n(n-1)$, то ε-равномерно устойчивый алгоритм оказывается абсолютно устойчивым.

Выписанные здесь характеристики абсолютной устойчивости можно уточнить в ряде важных специальных случаев. В частности, в некоторых кластерных алгоритмах (например, в алгоритме k-средних или неполном дивизимном алгоритме, подробно описанных в монографии Айвазяна С.А., Бухштабера В.М., Енюкова И.С., Мешалкина Л.Д. Прикладная статистика: Классификация и снижение размерности. – М., 1989), число множеств, составляющих начальное разбиение, не меняется в процессе работы. Для такого рода алгоритмов согласно теореме 4 ε-равномерная устойчивость превращается в абсолютную устойчивость уже при $\varepsilon = 3$.

Можно также ставить и решать задачи устойчивости кластерных алгоритмов в зависимости от степени плотности расположения тех объектов, которые мы разбиваем на кластеры. С точки зрения автора, вводимое здесь понятие устойчивости весьма содержательно и требует глубокого изучения.