Двухкомпонентная функция качества кластеризации множества элементов, представленных парными сравнениями*

С. Д. Двоенко
dsd@tsu.tula.ru
Тульский государственный университет, Россия, Тула, пр. Ленина, 92

Рассмотрены варианты известного алгоритма k-средних, в которых не требуется вычислять собственно средние по кластерам. В новых версиях алгоритма k-средних выполняются перестановки на матрице парных сравнений так, что в случае помещения анализируемого множества объектов в признаковое пространство достигается тот же самый результат кластеризации. Рассмотрена новая двухкомпонентная целевая функция качества кластеризации как минимизируемая комбинация внутрикластерных дисперсий (квадратов расстояний) с близостью кластеров между собой или, в двойственной формулировке, как максимизируемая комбинация внутрикластерных близостей с дисперсией (квадратами расстояний) между кластерами. Показано, что качество кластеризации удается улучшить по сравнению с обычным критерием качества кластеризации.

Ключевые слова: кластер; k-средних; расстояние; близость; беспризнаковый

Bi-partial objective function for clustering a set of elements in terms of pairwise comparisons*

S. D. Dvoenko
Tula State University, Russia, Tula, Lenin Ave., 92

Background: In a featureless case, a set of objects is represented only by results of pairwise mutual comparisons in the form of a distance, similarity, or kernel-based matrix. Nevertheless, the cluster centers can be implicitly represented by its distances to other objects without the feature space itself.

Methods: The present author proposes k-means clustering without computations of cluster centers at all. This novel procedure, referred to as the k-meanless clustering, makes permutations on the similarity or distance square matrix resulting in the same clustering for both featureless and feature-based cases. In addition, new bi-partial objective function combines intracluster distances with intercluster similarities and needs to be minimized or in the dual form combines intracluster similarities with intercluster distances and needs to be maximized.

Results: Based on bi-partial approach, the clustering quality can be improved relative to the usual objective function.

Concluding Remarks: The k-means idea is very popular in the form of many heuristic aggregating procedures where cluster centers cannot be explicitly presented. Therefore, they are only suboptimal versions of the k-means. The proposed k-meanless clustering is the correct version of them.

Keywords: cluster; k-means; distance; similarity; featureless

*Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00010.

Погружение элементов множества в метрическое пространство

В задаче кластер-анализа объекты $\omega_i \in \Omega$, $i = 1, \ldots, N$ обычно представлены как векторы $\mathbf{x}_i = (x_{i1}, \ldots x_{in})^{\mathrm{T}}$ в n-мерном пространстве признаков и образуют матрицу данных $X(N, n)$. В соответствии с гипотезой компактности объекты образуют локальные сгущения в виде K кластеров (классов, таксонов).

Хорошо известные алгоритмы типа k-средних [1] основаны на идее несмещенного разбиения [2]. В соответствии с ней каждый кластер Ω_k, $k = 1, \ldots, K$, представлен своим «представителем» $\tilde{\mathbf{x}}_k$, а центр кластера представлен средним $\bar{\mathbf{x}}_k$.

Если окажется, что для всех кластеров представители и центры совпадают $\tilde{\mathbf{x}}_k = \bar{\mathbf{x}}_k$, то получена несмещенная кластеризация, а противном случае – смещенная. Тогда необходимо назначить центры (средние объекты) в качестве новых представителей, заново расклассифицировать объекты по минимуму расстояния до представителей и вычислить новые центры кластеров.

В случае, когда признаковое пространство нам недоступно, средний объект $\omega(\bar{\mathbf{x}}_k)$ не представлен в матрице расстояний $D(N, N)$ как центр соответствующего кластера. Поэтому обычно в качестве эвристических агрегирующих процедур применяют некорректные версии алгоритма k-средних, где вместо центра кластера $\bar{\omega}_k$ в таком качестве используют объект, ближайший ко всем остальным в кластере. Тогда в общем случае при выполнении всех условий $\tilde{\omega}_k = \bar{\omega}_k$ может быть получена смещенная кластеризация, так как при погружении данного множества в соответствующее признаковое пространство окажется, что центр кластера $\mathbf{x}(\bar{\omega}_k)$ может не совпадать со средним объектом $\bar{\mathbf{x}}_k$.

Кластеризация по расстояниям до центров кластеров

Как известно, среднее арифметическое, используемое в качестве центра кластера, минимизирует его дисперсию и как результат дисперсию всей кластеризации [1].

Дисперсия кластера представлена квадратами отклонений объектов от центра кластера, т. е. квадратами соответствующих расстояний:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\mathbf{x}_i,\ \bar{\mathbf{x}}_k)\,.$$

Очевидно, что данный критерий минимизирует среднее квадратов расстояний до центра кластера и средневзвешенную дисперсию кластеризации в целом:

$$J(K) = \frac{1}{N} \sum_{k=1}^{K} N_k \sigma_k^2 = \sum_{k=1}^{K} \frac{N_k}{N} \sigma_k^2\,.$$

В отсутствие признаков средние объекты $\bar{\omega}_k$ обеспечивают несмещенную кластеризацию, также минимизируя дисперсии кластеров:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\omega_i,\ \bar{\omega}_k)$$

и значение критерия $J(K)$ в целом.

Если множество Ω будет помещено в соответствующее пространство признаков, где объекты $\mathbf{x}(\bar{\omega}_k)$ и $\bar{\mathbf{x}}_k$ совпадут, то два критерия:

$$J^X(K) = \min_{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_K} J(K);\quad J^D(K) = \min_{\bar{\omega}_1, \ldots, \bar{\omega}_K} J(K)$$

окажутся одинаковыми $J^X(K) = J^D(K)$. Очевидно, что $J^D(K) \geqslant J^X(K)$ в общем случае. Построим алгоритм для получения несмещенной кластеризации.

Для некоторого элемента $\omega_l \in \Omega$, взятого как начало координат, и пары $\omega_i,\ \omega_j$ их скалярное произведение $c_{ij} = (d_{li}^2 + d_{lj}^2 - d_{ij}^2)/2$ вычисляется на основе расстояний $d_{pq} = d(\omega_p,\ \omega_q)$, где $c_{ii} = d_{li}^2$ при $i = j$.

Следовательно, элементы главной диагонали матрицы $C_l(N, N)$ представляют собой квадраты расстояний от начала координат $\omega_l \in \Omega$ до остальных объектов. Удобно [3] поместить начало координат в центр тяжести множества $\omega_i \in \Omega,\ \ i = 1, \ldots, N$.

Как показано в [4, 5, 6, 7], можно немедленно доказать, что центр кластера $\bar{\omega}_k$ будет представлен своими расстояниями до остальных объектов $\omega_i \in \Omega,\ \ i = 1, \ldots, N$ без необходимости восстановления неизвестного нам признакового пространства, где N_k — число объектов в кластере Ω_k:

$$d^2(\omega_i,\ \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2;\ \omega_p,\ \omega_q \in \Omega_k\ ,$$

где дисперсия кластера вычисляется как:

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) = \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\ . \quad (1)$$

Известно, что алгоритм k-средних может быть представлен в различных вариантах в соответствии со способами пересчета средних в признаковом пространстве. Представим данный алгоритм для расстояний в нужном нам виде, где пересчет центров выполняется сразу после очередного переноса.

Алгоритм 1:

Шаг 0. Взять в качестве центров $\bar{\omega}_k^0,\ \ k = 1, \ldots, K$, например, K наиболее удаленных друг от друга объектов и назначить их представителями $\tilde{\omega}_k^0,\ \ k = 1, \ldots, K$.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $d(\omega_i,\ \bar{\omega}_k^s) \leqslant d(\omega_i,\ \bar{\omega}_j^s)$, где $j = 1, \ldots, K,\ j \neq k$.

2. Пересчитать, если требуется, центры $\bar{\omega}_k^s,\ \ k = 1, \ldots, K$ и представить их своими расстояниями до всех объектов $d(\omega_i,\ \bar{\omega}_k^s),\ \ i = 1, \ldots, N$.

3. Переместить следующий $i = i + 1$ объект ω_i.

4. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация, где $\tilde{\omega}_k^s = \bar{\omega}_k^s$, $k = 1, \ldots, K$, иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$ и перейти к следующему шагу $s = s + 1$.

Кластеризация перестановками без центров кластеров

Заметим, что можно также вычислить среднее квадратов расстояний между объектами в кластере. С учетом расстояний до себя получим выражение:

$$\eta_k' = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(\mathbf{x}_i - \mathbf{x}_j)^2 = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\mathbf{x}_i,\ \mathbf{x}_j)\ .$$

Из (1) для σ_k^2 немедленно следует, что $\eta'_k = 2\sigma_k^2$. Введем обозначение $\eta_k = \eta'_k/2 = \sigma_k^2$, где

$$\eta_k = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\omega_i,\ \omega_j)\ .$$

Следовательно, для всех кластеров минимизация взвешенных квадратов расстояний между объектами в кластерах приводит к минимизации средневзвешенной дисперсии кластеризации в целом:

$$\tilde{J}(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\eta_k = \sum_{k=1}^{K}\frac{N_k}{N}\eta_k\ .$$

Следовательно, критерии $\tilde{J}(K)$ и $J(K)$ совпадают $\tilde{J}(K) = J(K)$. Если множество Ω будет помещено в соответствующее пространство признаков, то объекты $\mathbf{x}(\bar{\omega}_k)$ и $\bar{\mathbf{x}}_k$ совпадут, где два критерия:

$$\tilde{J}^X(K) = \min_{\Omega_1,\dots,\Omega_K\in X}\tilde{J}(K) \text{ и } \tilde{J}^D(K) = \min_{\Omega_1,\dots,\Omega_K\in D}\tilde{J}(K)$$

также совпадут $\tilde{J}^X(K) = \tilde{J}^D(K)$. Очевидно, что в общем случае $\tilde{J}^D(K) \geqslant \tilde{J}^X(K)$.

Построим кластеризацию без центров. Очевидно, что такая кластеризация должна быть несмещенной, если для нее вычислить центры кластеров. Эквивалентная модификация алгоритма k-средних, рассмотренного выше, имеет следующий вид.

Алгоритм 2:

Шаг 0. Взять в качестве подмножеств $\Omega_k^0,\quad k = 1,\dots,K$, например, K наиболее компактных в некотором смысле подмножеств.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$ и принять $\tilde{J}^s(K) = \tilde{J}_k^s(K)$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$, выполнено условие $\tilde{J}_k^s(K) < \tilde{J}_j^s(K),\ j = 1,\dots K,\ j \neq k$.

2. Переместить следующий $i = i + 1$ объект ω_i.

3. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация. Иначе перейти к следующему шагу $s = s + 1$.

Кластеризация по близостям

Положительно полуопределенная матрица близостей $S(N,N)$ с элементами $s_{ij} = s(\omega_i,\ \omega_j) \geqslant 0$ может рассматриваться как матрица скалярных произведений в метрическом пространстве размерности не выше N. Относительно некоторой точки $\omega_k \in \Omega$, взятой как начало координат, где $s_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2$, $s_{ii} = d_{ki}^2$, расстояния определяются как $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$.

В данном случае центр кластера $\bar{\omega}_k$ может быть представлен своими близостями к остальным объектам $\omega_i \in \Omega,\ \ i = 1,\dots N$, где N_k – число объектов в кластере Ω_k:

$$s(\omega_i,\bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} s_{ip};\ \omega_p \in \Omega_k,\ \omega_i \in \Omega,\ \ i = 1,\dots,N\ .$$

Компактность кластера может быть представлена как средняя близость центра к остальным объектам в кластере:

$$\delta_k = \frac{1}{N_k}\sum_{i=1}^{N_k} s(\omega_i,\bar{\omega}_k) = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{p=1}^{N_k} s_{ip};\ \omega_i,\omega_p \in \Omega_k\ .$$

Несмещенная кластеризация минимизирует дисперсии кластеров σ_k^2 и максимизирует их компактности δ_k, где с учетом $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$ получим:

$$\sigma_k^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(s_{ii} + s_{jj} - 2s_{ij}) = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} s_{ij} = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k \ .$$

Тогда для всех кластеров получим:

$$J(K) = \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k\right) = \frac{1}{N}\sum_{i=1}^{N} s_{ii} - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k = c - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k \ .$$

Обозначим средневзвешенную компактность кластеризации, которую следует максимизировать, в виде нового функционала:

$$I(K) = \sum_{k=1}^{K}\frac{N_k}{N}\delta_k, \text{ где } \quad I(K) = c - J(K). \tag{2}$$

Немедленно получим две модификации алгоритма k-средних для близостей: с вычислением центров кластеров и без них. Построим алгоритм кластеризации с вычислением центров.

Алгоритм 3:

Шаг 0. Взять в качестве центров $\bar{\omega}_k^0, \quad k = 1, \ldots K$, например, K наименее близких друг к другу объектов и назначить их представителями $\tilde{\omega}_k^0, \ \ k = 1, \ldots, K$.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $s(\omega_i, \ \bar{\omega}_k^s) \geqslant s(\omega_i, \ \bar{\omega}_j^s)$, где $j = 1, \ldots, K, \ j \neq k$.

2. Пересчитать, если требуется, центры $\bar{\omega}_k^s, \ k = 1, \ldots K$ и представить их своими близостями ко всем объектам $s(\omega_i, \ \bar{\omega}_k^s), \ \ i = 1, \ldots, N$.

3. Переместить следующий $i = i + 1$ объект ω_i.

4. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация, где $\tilde{\omega}_k^s = \bar{\omega}_k^s, \ k = 1, \ldots, K$, иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$ и перейти к следующему шагу $s = s + 1$.

Построим алгоритм кластеризации по близостям без центров.

Алгоритм 4:

Шаг 0. Взять в качестве подмножеств $\Omega_k^0, \quad k = 1, \ldots, K$, например, K наиболее компактных в некотором смысле подмножеств.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$ и принять $I^s(K) = I_k^s(K)$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$, выполнено условие $I_k^s(K) > I_j^s(K), \ j = 1, \ldots, K, \ j \neq k$.

2. Переместить следующий $i = i + 1$ объект ω_i.

3. Стоп, если ни один объект не был перемещен в другой кластер, т.е. получена несмещенная кластеризация. Иначе перейти к следующему шагу $s = s + 1$.

Двухкомпонентная целевая функция качества кластеризации

Легко увидеть, что для всех вариантов классического критерия качества кластеризации как для признакового пространства, так и в случае только парных сравнений, справедливо общее свойство: минимизируя разброс объектов в кластерах (или максимизируя

«плотность» кластеров в двойственной формулировке) мы никак не управляем разбросом центров кластеров. Очевидно, что в общем случае, делая кластеры более плотными, желательно еще попробовать и отдалить их друг от друга, насколько это возможно.

Реализация такого подхода [8, 9] приводит к построению двухкомпонентной целевой функции качества кластеризации. При построении такой функции возникает проблема масштабирования двух ее частей: одна из них отвечает за внутриклассовые характеристики, а другая – за межклассовые. Это приводит в общем случае к необходимости выбора соответствующих шкал измерений, зависящих от интерпретации понятия «кластер» и к согласованию их путем поиска оптимальной линейной комбинации.

В нашем случае, в отличие от рассмотренного подхода, задача оказывается проще. А именно: нам не требуется поиск согласованных шкал измерений для внутри- и межклассовых характеристик качества кластеризации, так как в беспризнаковом подходе расстояния и близости между элементами множества представлены в одном и том же метрическом пространстве, пусть даже и неизвестном. В этом случае потребуется просто определить масштаб влияния дополнительной части на общее значение критерия качества путем подбора соответствующего коэффициента.

Рассмотрим снова критерий $J(K)$ и его вариант $\tilde{J}(K)$ при отсутствии явно вычисленных центров кластеров. С учетом (1) для дисперсии η_k кластера Ω_k получим:

$$\tilde{J}(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\eta_k = \frac{1}{N}\sum_{k=1}^{K}\frac{N_k}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 = \frac{1}{2N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 \ .$$

Согласно (2), в двойственной формулировке для критерия $I(K)$ также получим:

$$I(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\delta_k = \frac{1}{N}\sum_{k=1}^{K}\frac{N_k}{N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} s_{pq} = \frac{1}{N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} s_{pq} \ .$$

Сначала рассмотрим двухкомпонентную целевую функцию качества кластеризации $\tilde{J}_\delta(K) = \tilde{J}(K) + \delta(K)$, которую нужно минимизировать, при комбинировании внутрикластерных дисперсий $\tilde{J}(K)$ с близостью между кластерами $\delta(K)$.

Как и ранее для кластеров, рассмотрим центр всего множества и обозначим его как новый элемент $\bar{\omega}_0$, который представим своими близостями в данном случае не ко всем элементам множества, а только к центрам других кластеров:

$$s(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K}\sum_{p=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_p),\ k = 1, \ldots, K \ .$$

Компактность множества центров кластеров, которую будем рассматривать как близость между кластерами $\delta(K)$, может быть представлена как средняя близость центра всего множества $\bar{\omega}_0$ к центрам кластеров $\bar{\omega}_k,\ k = 1, \ldots, K$:

$$\delta(K) = \frac{1}{K}\sum_{k=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_0) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{K}\sum_{l=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_l) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} s(\bar{\omega}_k,\ \bar{\omega}_l) \ .$$

Определим близости $s(\bar{\omega}_k,\ \bar{\omega}_l)$ центров кластеров друг к другу. Центр кластера $\bar{\omega}_k$ представлен своими близостями ко всем остальным объектам $\omega_i \in \Omega$ и, в частности, к объектам из другого кластера $\omega_i \in \Omega_l$. Рассмотрим среднюю близость объектов из другого

кластера $\omega_i \in \Omega_l$ к центру данного кластера $\bar{\omega}_k$:

$$s(\Omega_l,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l} s(\omega_i,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l}\frac{1}{N_k}\sum_{p=1}^{N_k} s_{ip} = \frac{1}{N_l N_k}\sum_{i=1}^{N_l}\sum_{p=1}^{N_k} s_{ip},\ \omega_p \in \Omega_k\ .$$

Очевидно, что $s(\Omega_l,\ \bar{\omega}_k) = s(\Omega_k,\ \bar{\omega}_l)$, так как $s_{ij} = s_{ji}$. Следовательно, можно использовать обозначения $s(\Omega_l,\ \bar{\omega}_k) = s(\Omega_k,\ \bar{\omega}_l) = s(\Omega_l,\ \Omega_k) = s(\bar{\omega}_l,\ \bar{\omega}_k)$ для парной близости между кластерами. В итоге близость между всеми кластерами выражается следующим образом:

$$\delta(K) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} s_{pq},\quad \text{где}\quad \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ .$$

Теперь рассмотрим двухкомпонентную целевую функцию качества кластеризации $I_\sigma(K) = I(K) + \sigma^2(K)$, которую нужно максимизировать, при комбинировании внутрикластерных близостей $I(K)$ с межкластерной дисперсией $\sigma^2(K)$.

Рассмотрим центр всего множества как объект $\bar{\omega}_0$, который представим своими расстояниями до центров других кластеров. Как показано в [4, 5, 6] и согласно (1) получим:

$$d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q),\ k = 1, \ldots, K\ .$$

Дисперсия множества центров кластеров $\sigma^2(K)$ может быть представлена как среднее квадратов расстояний от центра всего множества $\bar{\omega}_0$ до центров кластеров $\bar{\omega}_k,\ k = 1, \ldots, K$:

$$\sigma^2(K) = \frac{1}{K}\sum_{k=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_0) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q)\right) =$$

$$= \frac{1}{K^2}\sum_{k=1}^{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k,\ \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q) = \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p,\ \bar{\omega}_q)\ .$$

Определим расстояния $d^2(\bar{\omega}_k,\ \bar{\omega}_l)$ между центрами кластеров. Центр кластера $\bar{\omega}_k$ представлен своими расстояниями до остальных объектов $\omega_i \in \Omega$ и, в частности, до объектов из другого кластера $\omega_i \in \Omega_l$. Рассмотрим среднее квадратов расстояний объектов из другого кластера $\omega_i \in \Omega_l$ до центра данного кластера $\bar{\omega}_k$:

$$d^2(\Omega_l,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l} d^2(\omega_i,\ \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) =$$

$$= \frac{1}{N_l N_k}\sum_{i=1}^{N_l}\sum_{p=1}^{N_k} d_{ip}^2\ - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2,\ \ \omega_p,\ \omega_q \in \Omega_k\ .$$

Аналогично получим:

$$d^2(\Omega_k,\ \bar{\omega}_l) = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\omega_i,\ \bar{\omega}_l) = \frac{1}{N_k N_l}\sum_{i=1}^{N_k}\sum_{p=1}^{N_l} d_{ip}^2\ - \frac{1}{2N_l^2}\sum_{p=1}^{N_l}\sum_{q=1}^{N_l} d_{pq}^2,\ \omega_p,\ \ \omega_q \in \Omega_l\ .$$

Легко увидеть, что в общем случае $d^2(\Omega_l,\ \bar{\omega}_k) \neq d^2(\Omega_k,\ \bar{\omega}_l)$ из-за различных внутрикластерных дисперсий кластеров Ω_l и Ω_k.

Рассмотрим величину $d^2(\Omega_l,\ \Omega_k) = (1/2)\,(d^2(\Omega_l,\ \bar{\omega}_k) + d^2(\Omega_k,\ \bar{\omega}_l))$ как расстояние между двумя множествами. Очевидно, что при $d_{ij} = d_{ji}$ получим:

$$d^2(\Omega_l,\ \Omega_k) = \frac{1}{N_k N_l}\sum_{i=1}^{N_k}\sum_{p=1}^{N_l} d_{ip}^2 - \frac{1}{2}\left(\frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 + \frac{1}{2N_l^2}\sum_{s=1}^{N_l}\sum_{t=1}^{N_l} d_{st}^2\right),$$

где $\omega_p,\ \omega_q \in \Omega_k$ и $\omega_s,\ \omega_t \in \Omega_l$. В этом случае можно также ввести обозначение:

$$d^2(\bar{\omega}_l,\ \bar{\omega}_k) = d^2(\Omega_l,\ \Omega_k) = \frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2 - \frac{1}{2}(\sigma_k^2 + \sigma_l^2)\ .$$

Тогда межкластерная дисперсия выражается следующим образом:

$$\sigma^2(K) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2 - \frac{1}{2}(\sigma_k^2 + \sigma_l^2)\right),\ \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ .$$

Если внутрикластерные дисперсии не учитывать, то окажется, что $d^2(\Omega_l,\ \bar{\omega}_k) = d^2(\Omega_k,\ \bar{\omega}_l)$, так как $d_{ij} = d_{ji}$. Тогда для парных расстояний между центрами кластеров также можно использовать обозначения $d^2(\Omega_l,\ \bar{\omega}_k) = d^2(\Omega_k,\ \bar{\omega}_l) = d^2(\Omega_l,\ \Omega_k) = d^2(\bar{\omega}_l,\ \bar{\omega}_k)$. В этом случае межкластерная дисперсия выражается следующим образом:

$$\sigma^2(K) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2\ ,\ \text{где}\quad \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ .$$

Очевидно, что при $K = N$ в обоих случаях мы получим дисперсию всего множества, так как дисперсии одноэлементных кластеров являются нулевыми.

Таким образом, в случае двухкомпонентной функции качества кластеризации применяются те же алгоритмы k-средних (варианты 2 и 4), но только для критериев $\tilde{J}_\delta(K)$ и $I_\sigma(K)$.

Вычисление компонент целевой функции

Очевидно, что в двухкомпонентных целевых функциях $\tilde{J}_\delta(K)$ и $I_\sigma(K)$ необходимо одновременно использовать представление элементов множества как расстояниями, так и близостями между ними. Пусть задана матрица близостей $S(N,N)$ с элементами $s_{ij} = s(\omega_i,\ \omega_j) \geqslant 0$. Тогда элементы матрицы расстояний $D(N,N)$ получаются преобразованием $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$.

Пусть теперь задана матрица расстояний $D(N,N)$. Чтобы получить близости, необходимо назначить начало координат как некоторый объект ω_0, относительно которого можно по теореме косинусов вычислить скалярные произведения $s_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$, где для ненормированных величин диагональные элементы $s_{ii} = d_{0i}^2$ представляют расстояния всех остальных объектов $\omega_i \in \Omega,\ \ i = 1, \ldots N$ до начала координат ω_0.

Начало координат ω_0 можно выбрать разными способами, например, по методу главных проекций Торгерсона [3] поместить его в центр тяжести множества Ω. Тогда объект ω_0 будет представлен своими расстояниями до остальных объектов следующим образом:

$$d_{0i}^2 = d^2(\omega_0,\ \omega_i) = \frac{1}{N}\sum_{p=1}^{N} d_{ip}^2 - \frac{1}{2N^2}\sum_{p=1}^{N}\sum_{q=1}^{N} d_{pq}^2;\ \omega_p,\ \omega_q \in \Omega\ . \qquad (3)$$

Тем не менее такое представление окажется неудобным, так как относительно такого начала координат скалярные произведения s_{ij} могут оказаться как положительными, так и отрицательными. Необходимо так назначить начало координат, чтобы все скалярные произведения объектов относительно него были бы неотрицательными $s_{ij} \geqslant 0$. Только в этом случае мы можем рассматривать полученные значения именно как близости между элементами множества в метрическом пространстве. Содержательно это означает, что относительно такого начала координат все объекты должны располагаться в положительном квадранте координатного пространства, т. е. такое новое начало координат должно располагаться вне выпуклой оболочки, образованной данным множеством объектов, и на некотором достаточном удалении от множества объектов.

Как и ранее, отметим, что второе слагаемое в (3) представляет собой дисперсию множества:

$$\sigma^2 = \frac{1}{2N^2} \sum_{p=1}^{N} \sum_{q=1}^{N} d_{pq}^2 \, .$$

Рассмотрим первое слагаемое в (3). Рассмотрим расстояния $d_{ip},\ p = 1, \ldots N$ от объекта ω_i до остальных объектов ω_p как компоненты соответствующего вектора в некотором N-мерном пространстве, которое удобно считать «вторичным». Тогда величина $\sum_{p=1}^{N} d_{ip}^2$ представляет собой квадрат нормы этого вектора, т.е. квадрат расстояния от начала координат, а первое слагаемое из (3) представляет собой среднее квадрата этой нормы.

Таким образом, начало координат в таком вторичном пространстве будет представлено как объект ω_{0^0} своими расстояниями до остальных объектов $d_{0^0 i}^2,\ i = 1, \ldots N$, инвариантными относительно размера множества. Тогда начало координат по методу Торгерсона будет представлено как объект ω_0 своими расстояниями $d_{0i}^2 = d_{0^0 i}^2 - \Delta$, где $\Delta = \sigma^2$.

Легко увидеть, что изменение константы Δ позволит получить начало координат не в центре тяжести множества. При $\Delta = 0$ начало координат как объект ω_0 будет максимально удалено от центра тяжести множества. Условие $\Delta = 0$ можно понимать как наименьший разброс элементов множества по сравнению с расстояниями до начала координат. Очевидно, что в этом случае скалярные произведения будут близки к единице, если получившиеся расстояния до начала координат окажутся значительными. Если $\Delta < 0$, то это свойство только усилится.

При $\Delta > \sigma^2$ обязательно возникнет ситуация, когда не удастся получить корректный вид матрицы скалярных произведений $S(N, N)$. Это произойдет, когда некоторые из расстояний $d_{0i}^2 = d_{0^0 i}^2 - \Delta$ до начала координат окажутся нулевыми или отрицательными. В этом случае условие $\Delta > \sigma^2$ можно понимать как увеличенный по сравнению с реальным разброс элементов множества. В итоге допустимые значения величины Δ находятся в интервале $0 \leqslant \Delta \leqslant \sigma^2$.

Эксперименты

Были проведены эксперименты на данных по ирисам [10], которые представляют собой измерения четырех признаков (длина и ширина чашелистика, длина и ширина лепестка) пятидесяти экземпляров растений каждого из трех видов (Iris Setosa — «касатик щетиноносный», Iris Versicolor — «касатик разноцветный», Iris Virginica — «касатик виргинский»), всего 150 экземпляров.

Известно, что первый класс (Iris Setosa) хорошо отделен от остальных двух классов (второй класс — Iris Versicolor, третий класс — Iris Virginica), которые слегка пересекаются между собой. Так как классификация объектов для этих данных заранее известна,

Таблица 1. Ирисы. Разделение классов

Нач. разбиение	Ошибки ($\alpha = 0$)	α_{opt}	Ошибки (α_{opt})
50-50-50	16	3–6	15
50-70-30	16	3–6	15
50-30-70	16	3–6	15
50-50	16	12–17.7	15
70-30	16	12–17.7	15
30-70	16	12–17,7, 22–22,4	15

то необходимо показать, что применение двухкомпонентной функции критерия качества кластеризации позволяет объективно улучшить результат разбиения, правильно отделив первый класс от остальных и уменьшив ошибки кластеризации для второго и третьего классов.

Для примера рассмотрим критерий $\tilde{J}_\delta(K) = \tilde{J}(K) + \alpha\delta(K)$, где нужно будет подобрать масштабирующий коэффициент α для дополнительной части.

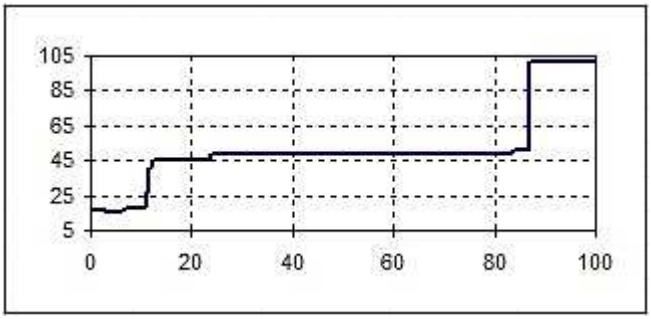
Известно, что алгоритм k-средних дает локально-оптимальный результат, который зависит от начального разбиения. Поэтому в экспериментах начальные разбиения фиксировались заранее, чтобы иметь возможность для сравнения результатов. Для трех классов использовались начальные разбиения: 50-50-50 (исходное), 50-70-30 (20 объектов третьего класса ошибочно отнесены ко второму классу) и 50-30-70 (20 объектов второго класса ошибочно отнесены к третьему классу). Для двух пересекающихся классов (второй и третий) использовались начальные разбиения: 50-50 (исходное), 70-30 (20 объектов третьего класса ошибочно отнесены ко второму классу) и 30-70 (20 объектов второго класса ошибочно отнесены к третьему классу).

В первой серии экспериментов было показано (см. табл. 1 и рис. 1), что для всех начальных разбиений удается подобрать оптимальное значение коэффициента α, при котором для случая трех классов первый отделяется безошибочно, а число ошибок разделения второго и третьего классов уменьшается по сравнению с классическим критерием ($\alpha = 0$). То же самое было показано и для случая двух пересекающихся классов (без первого). В обоих экспериментах ошибочными оказались одни и те же объекты: 102, 107, 114, 115, 120, 122, 124, 127, 128, 134, 135, 139, 143, 147, 150. Также видно, что оптимальное значение коэффициента α зависит от соотношения дисперсий разделяемых множеств.

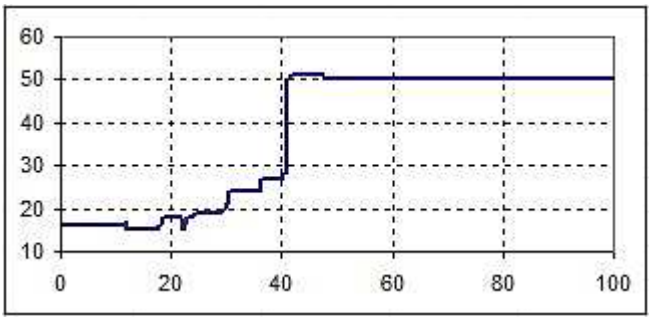
Во второй серии экспериментов рассматривалась известная проблема, когда разделение резко различающихся по размеру кластеров приводит к ошибкам, так как критерий $\tilde{J}(K)$ стремится разбить большой кластер, образованный вторым и третьем классами, и увеличить размер небольшого кластера, образованного первым классом. Таким образом, при $\alpha = 0$ три объекта 58, 94 и 99 были неправильно отнесены к первому классу.

Здесь также удается подобрать оптимальное значение коэффициента $\alpha = 27$, при котором достигается безошибочное разделение. В данном случае небольшой первый класс безошибочно отделился от второго и третьего классов, рассмотренных вместе как большой класс (рис. 1).

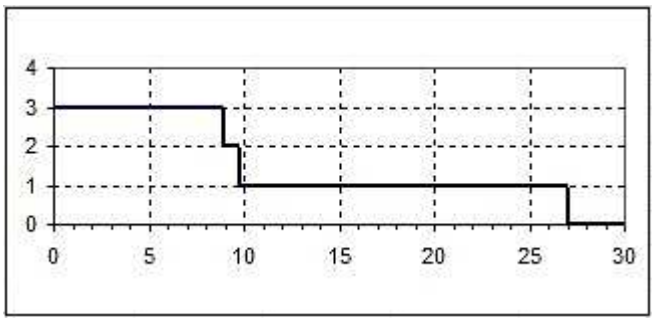
В случае, когда отсутствует признаковое пространство, множество элементов представлено только результатами их парных сравнений в виде матрицы близостей или расстояний.



Setosa – Versicolor – Virginica (50-50-50, 50-70-30, 50-30-70)



Versicolor – Virginica (30-70)



Setosa – Versicolor&Virginica (50-100, 100-50, 30-120)

Рис. 1. Ирисы. Графики числа ошибок при подборе оптимального значения α

В данной работе рассмотрены новые версии известного алгоритма k-средних для случая, когда пространство исходных признаков неизвестно. Иногда оказывается, что в этом случае удобнее не вводить понятие среднего объекта, так как возможны трудности с его интерпретацией, а применять перестановочные версии агрегирующих процедур.

В данной работе построены корректные версии таких процедур, которые дают одинаковый результат, если полученное разбиение окажется все-таки помещенным в подходящее признаковое пространство.

Применение двухкомпонентной целевой функции позволяет улучшить качество кластеризации, что было показано на известных данных по ирисам.

Очевидно, что в таких перестановочных алгоритмах необходимо снижать их сложность, выполняя пересчет значений критерия на основе приращений при переносах объектов между кластерами. Это можно делать экономно на основе рекуррентных соотношений, и такие способы известны.

Литература

[1] *Duda R. O., Hart P. E., Stork D. G.* Pattern classification. N.Y.: Wiley, 2001. 654 p.

[2] *Шлезингер М. И.* О самопроизвольном различении образов // *Читающие автоматы и распознавание образов.* Киев: Наукова думка, 1965. С. 38–45.

[3] *Torgerson W. S.* Theory and methods of scaling. N.Y.: Wiley, 1958. 460 p.

[4] *Двоенко С. Д.* Кластеризация элементов множества на основе взаимных расстояний и близостей // *ММРО-13.* М: МАКС-Пресс, 2007. С. 114–117.

[5] *Двоенко С. Д.* Кластеризация множества, описанного парными близостями и расстояниями между его элементами // *Сибирский журнал индустриальной математики*, 2009. Т. 12, № 1. С. 61–73.

[6] *Dvoenko S. D.* Clustering and separating of a set of members in terms of mutual distances and similarities // *Trans. Machine Learning Data Mining*, 2009. Vol. 2, No. 2. P. 80–99.

[7] *Dvoenko S. D.* On featureless k-means clustering // *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*, 2013. Tilburg, the Netherlands. P. 70.

[8] *Owsinski J. W.* The bi-partial approach in clustering and ordering: the model and the algorithms // *Statistica & Applicazioni. Special Issue*, 2011. P. 43–59.

[9] *Owsinski J. W.* The matter of scale: Perceiving distances and proximities in the bi-partial clustering setting // *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*, 2013. Tilburg, the Nethelands. P. 88–89.

[10] *Fisher R. A.* The use of multiple measurements in taxonomic problems // *Ann. Eugenics*, 1936. Vol. 7, No. 9. P. 179–188.

References

[1] *Duda R. O., Hart P. E., Stork D. G.* 2001. Pattern classification. N.Y.: Wiley. 654 p.

[2] *Schlesinger M. I.* 1965. On spontaneous pattern distinguishing. *Reading Automata and Pattern Recognition.* Kiev: Naukova Dumka. 38–45. (in Russ.)

[3] *Torgerson W. S.* 1958. Theory and methods of scaling. N.Y.: Wiley. 460 p.

[4] *Dvoenko S. D.* 2007. Clustering a set members by mutual distances ans similarities. *MMPR-13.* M: MAKS-Press. 114–117. (in Russ.)

[5] *Dvoenko S. D.* 2009. Clusterization of the set presented by distances and similarities between its elements. *Syberian Journal of Industrial Mathematics* 12(1):61–73. (in Russ.)

[6] *Dvoenko S. D.* 2009. Clustering and separating of a set of members in terms of mutual distances and similarities. *Trans. Machine Learning Data Mining* 2(2):80–99.

[7] *Dvoenko S. D.* 2013. On featureless k-means clustering. *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*. Tilburg, the Nethelands. 70.

[8] *Owsinski J. W.* 2011. The bi-partial approach in clustering and ordering: the model and the algorithms. *Statistica & Applicazioni. Special Issue*. 43–59.

[9] *Owsinski J. W.* 2013. The matter of scale: Perceiving distances and proximities in the bi-partial clustering setting. *Conference of the International Federation of Classification Societies (IFCS-2013) Abstracts*. Tilburg, the Netherlands. 88–89.

[10] *Fisher R. A.* 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7(9):179–188.