Math-Net.Ru

Общероссийский математический портал

А. А. Болотов, О метрической кластеризации, *Дискрет. матем.*, 1996, том 8, выпуск 4, 62–78

DOI: <https://doi.org/10.4213/dm543>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 54.174.149.18

26 апреля 2023 г., 15:40:47

О метрической кластеризации

© 1996 г. А. А. Болотов

В рамках комбинаторно-логического подхода изучается задача кластеризации векторов, заданных в n -мерном пространстве признаков. Основными результатами работы являются изучение свойств оптимальных относительно рассматриваемых функционалов разбиений на кластеры и построение возникающих на этой основе итерационных и градиентных алгоритмов кластеризации. Показано, что все оптимальные разбиения являются хорошо разделимыми относительно ассоциированных решающих правил.

1. Введение

Комбинаторно-логический подход к задачам распознавания, прогнозирования и классификации оформился в новое направление в теории распознавания образов в 70-е годы в работах В. Б. Кудрявцева и его учеников. В этот и последующий периоды времени здесь было получено много интересных результатов, как теоретических, так и имеющих важное практическое значение.

В настоящей работе в рамках комбинаторно-логического подхода изучается задача кластеризации векторов, заданных в n -мерном пространстве Π признаков, снабженном структурой линейного векторного пространства,

$$\Pi = \Pi_1 \times \dots \times \Pi_n,$$

где Π_i , $i = 1, \dots, n$, определяет шкалу измерения i -го признака заданных объектов. В результате измерения всех признаков объектов $O_1, \dots, O_N$ получаем N векторов $x_1, \dots, x_N$ из пространства Π , которые необходимо разбить на кластеры (группы) в соответствии с требованием о компактности векторов, образующих тот или иной кластер, и различии кластеров между собой. Разбиение данного множества векторов $X = \{x_1, \dots, x_N\}$ на q непустых кластеров

$$K_1, \dots, K_q, \quad \bigcup_i K_i = X, \quad K_i \cap K_j = \emptyset, \quad i \neq j,$$

будем обозначать $\Sigma_q(X)$, соответствующее разбиение множества индексов $\{1, 2, \dots, N\}$ на q групп $k_1, \dots, k_q$ обозначаем $\sigma_q(N)$, так что

$$K_i = \{x_j \mid j \in k_i\}, \quad i = 1, \dots, q,$$

при фиксированных X и N будем писать просто Σ_q и σ_q .

Поставленная задача уточняется в работе следующим образом. Вещественнозначную функцию $\Phi(x, y)$ называем потенциалом поля (физические термины используются по аналогии и для удобства), создаваемого точкой $x \in \Pi$ в точке $y \in \Pi$, если она является

- (1) центрально-симметричной (относительно точки x), т.е.

$$\Phi(x, y) = \langle x - y \rangle,$$

где $\langle \cdot \rangle: \Pi \rightarrow R$ — некоторый функционал со свойствами

$$(a) \langle x \rangle \geq 0, x \in \Pi; \langle x \rangle = 0 \iff x = 0,$$

$$(б) \langle x \rangle = \langle -x \rangle;$$

- (2) аддитивной, т.е. потенциал $\Phi(x, y)$, создаваемый в точке y множеством точек K , вычисляется по формуле

$$\Phi(K, y) = \sum_{x \in K} \Phi(x, y) = \sum_{x \in K} \langle x - y \rangle.$$

Потенциал Φ такого типа будем называть метрическим потенциалом. Используя метрический потенциал, можно определять различные характеристики кластера (например, энергию кластера, потенциал кластера, размер кластера и т. д.) и, соответственно, аналогичные интегральные характеристики разбиения на кластеры, что позволяет формально уточнить задачу кластеризации как оптимизационную задачу определения разбиения с оптимальными характеристиками.

Во втором параграфе этой работы изучаются свойства оптимальных разбиений относительно энергии, в третьем, при дополнительном свойстве метрического потенциала, относительно потенциала и в четвертом параграфе относительно размера. В техническом отношении настоящая работа примыкает к статье [1].

2. Кластеризация с минимальной энергией

Пусть Σ_q — разбиение векторов на q кластеров $K_1, \dots, K_q$ и σ_q — соответствующее разбиение множества индексов векторов на соответствующие множества $k_1, \dots, k_q$. Обозначим $m_j = |K_j|$, так что $m_1 + \dots + m_q = N$.

Для кластера K_s определим его потенциальную энергию P_s , полагая

$$P_s = \frac{1}{2} \sum_{i \in k_s} \sum_{j \in k_s} \langle x_i - x_j \rangle = \sum_{i < j} \langle x_i - x_j \rangle. \quad (1)$$

Для двух кластеров K_r и K_s потенциальной энергией этих кластеров относительно друг друга называем величину

$$P(K_r, K_s) = \sum_{i \in k_r} \sum_{j \in k_s} \langle x_i - x_j \rangle. \quad (2)$$

Отметим, что для одноточечного кластера $K_s = \{y\}$ величина $P(K_r, \{y\})$ есть потенциал $\Phi(K_r, y)$, создаваемый кластером K_r в точке y . Далее в подобных обозначениях будем опускать фигурные скобки для одноэлементных множеств. Отметим также, что $P(K_s, K_s) = 2P_s$.

Из свойства аддитивности потенциала и определения (2) вытекает свойство аддитивности потенциальной энергии взаимодействия кластеров:

$$\begin{aligned} P(K_p \cup K_r, K_s) &= P(K_p, K_s) + P(K_r, K_s), \\ P(K_p \setminus K_r, K_s) &= P(K_p, K_s) - P(K_r, K_s), \quad K_r \subseteq K_p. \end{aligned} \quad (3)$$

Вычислим, пользуясь свойством аддитивности (3), потенциальную энергию P всей совокупности векторов:

$$\begin{aligned} P &= \frac{1}{2}P(X, X) = \frac{1}{2}P(K_1 \cup \dots \cup K_q, K_1 \cup \dots \cup K_q) \\ &= \sum_{i=1}^q P_i + \sum_{i < j} P(K_i, K_j). \end{aligned} \quad (4)$$

Первое слагаемое в правой части равенства (4) характеризует компактность кластеров, в то время как второе слагаемое — степень различия между кластерами. Так как левая часть равенства (4) не зависит от разбиения, то минимизируя слагаемое $\sum_{j=1}^q P_j$, мы одновременно максимизируем слагаемое $\sum_{i < j} P(K_i, K_j)$.

Величину

$$P(\Sigma_q) = \sum_{j=1}^q P_j$$

будем называть потенциальной энергией разбиения Σ_q , а разбиение $\bar{\Sigma}_q$ с минимальным значением энергии (при фиксированном q), т.е. $P(\bar{\Sigma}_q) = \min P(\Sigma_q)$, где минимум берется по всем разбиениям на q кластеров, оптимальным (относительно энергии).

Легко видеть, что потенциальная энергия оптимального разбиения как функция от q есть строго убывающая функция. Действительно, при $q = N$ справедливо равенство $P(\bar{\Sigma}_N) = 0$. Далее, при $q \leq N - 1$ некоторый кластер K_i в оптимальном разбиении $\bar{\Sigma}_q = \{K_1, \dots, K_q\}$ содержит, по крайней мере, два элемента x и y , $x, y \in K_i$. Образует новое разбиение на $q + 1$ кластер

$$\Sigma_{q+1} = \{K_1, \dots, K_i \setminus \{x\}, \dots, K_q, \{x\}\}.$$

Очевидно,

$$P(\bar{\Sigma}_q) > P(\Sigma_{q+1}), \quad P(\Sigma_{q+1}) \geq P(\bar{\Sigma}_{q+1}),$$

так как Σ_{q+1} может и не быть оптимальным разбиением. В итоге получаем, что

$$P(\bar{\Sigma}_q) > P(\bar{\Sigma}_{q+1}).$$

Нам понадобятся формулы, выражающие энергию кластера, полученного слиянием (или разбиением) двух других. Пусть, например, кластер K_l образован слиянием двух кластеров K_i и K_j , так что $K_l = K_i \cup K_j$. Тогда

$$\begin{aligned} P_l &= \frac{1}{2}P(K_l, K_l) = \frac{1}{2}P(K_i \cup K_j, K_i \cup K_j) = \frac{1}{2}P(K_i, K_i) + \frac{1}{2}P(K_j, K_j) + P(K_i, K_j) \\ &= P_i + P_j + P(K_i, K_j). \end{aligned} \quad (5)$$

Равенство (5) выражает (в определенном смысле) закон сохранения энергии при слиянии кластеров. В частности, если $K_i = \{x\}$ — одноточечный кластер, то из (5) следует, что

$$P'_j = P_j + \Phi(K_j, x), \quad (6)$$

где P'_j обозначает энергию кластера $K_j \cup \{x\}$.

В случае разбиения кластера на два, $K_l = K_i \setminus K_j$, $K_j \subseteq K_i$, используя (5) и (3), получим, что

$$P_l = P_i - (P(K_i, K_j) - P_j). \quad (7)$$

В частности, если $K_j = \{x\}$ — одноточечный кластер, то из (7) следует, что

$$P'_i = P_i - \Phi(K_i, x), \quad (8)$$

где P'_i обозначает энергию кластера $K_i \setminus \{x\}$, $x \in K_i$.

В рамках энергетического подхода естественным образом возникает следующее решающее правило

$$\mathcal{R}_1: x \in K_i \iff \Phi(K_i, x) < \Phi(K_j, x), \quad j \neq i,$$

где $\in$ означает относится к классу. В случае, когда минимум выражения $\Phi(K, x)$ достигается на двух или более классах, происходит отказ от классификации.

Разбиение $\Sigma_q = \{K_1, \dots, K_q\}$ будем называть $\mathcal{R}$ -разделимым относительно решающего правила $\mathcal{R}$, если для всех $x \in X$ и для всех $i = 1, \dots, q$

$$x \in K_i \iff x \in K_i, \quad (9)$$

т.е. решающее правило $\mathcal{R}$ правильно классифицирует векторы из всех классов. Если (9) выполняется с точностью до векторов, на которых происходит отказ от классификации, то будем говорить о слабой $\mathcal{R}$ -разделимости разбиения Σ_q . Справедливо следующее утверждение.

Теорема 1. *Оптимальные разбиения $\bar{\Sigma}_q$ являются слабо $\mathcal{R}_1$ -разделимыми.*

Доказательство. Пусть $\bar{\Sigma}_q$ — оптимальное разбиение. Очевидно, что решающее правило $\mathcal{R}_1$ правильно классифицирует векторы из одноэлементных классов. Пусть некоторый класс K_i состоит из более, чем одного вектора, и пусть x — произвольный вектор из K_i . Выберем произвольный класс K_j , $j \neq i$, и рассмотрим разбиение Σ_q , полученное из $\bar{\Sigma}_q$ в результате переноса вектора x из класса K_i в класс K_j , т.е.

$$\Sigma_q = \{K_1, \dots, K_i \setminus \{x\}, \dots, K_j \cup \{x\}, \dots, K_q\}.$$

Для энергии классов $K'_i = K_i \setminus \{x\}$ и $K'_j = K_j \cup \{x\}$ из (6) и (8) получим, что

$$P'_i = P_i - \Phi(K_i, x), \quad P'_j = P_j + \Phi(K_j, x).$$

Тогда изменение потенциальной энергии ΔP составит

$$\Delta P = P(\Sigma_q) - P(\bar{\Sigma}_q) = \Phi(K_j, x) - \Phi(K_i, x) \geq 0,$$

так как $\bar{\Sigma}_q$ — оптимальное разбиение.

Таким образом, $\Phi(K_i, x) \leq \Phi(K_j, x)$, что и означает (при отсутствии отказа от классификации в случае равенства) правильную классификацию. Теорема доказана.

Теорема 1 дает необходимое условие оптимальности разбиения Σ_q , по крайней мере, оно должно быть слабо $\mathcal{R}_1$ -разделимым. На этом основывается итерационная процедура нахождения разбиения близкого к оптимальному: осуществляется проверка условия (9) и при его нарушении (со строгим неравенством) вектор x переносится из класса K_i в класс K_j . Можно показать, что такая процедура за конечное число итераций (в случае конечного множества X) сходится к некоторому слабо $\mathcal{R}_1$ -разделимому разбиению.

Принципиальные трудности при построении оптимальных разбиений $\bar{\Sigma}_q$ связаны с тем, что соответствующая целочисленная задача кластеризации с минимальной энергией, состоящая в том, чтобы при заданных конечном множестве X , значениях $\langle x_i - x_j \rangle \in \mathbb{Z}_0^+$ для каждой пары $x_i, x_j \in X$ и двух натуральных числах q и P , определить существует ли разбиение $\Sigma_q(X)$ с энергией

$$P(\Sigma_q) \leq P,$$

является NP-полной. Действительно, уже для $q = 2$, принимая во внимание (4), можно к сформулированной задаче полиномиально свести известную NP-полную задачу теории графов о минимальном разрезе.

В связи с этим представляют интерес приемлемые приближенные процедуры построения разбиений, близких к оптимальным.

Будем максимизировать величину

$$R(\Sigma_q) = \sum_{i < j} P(K_i, K_j),$$

что эквивалентно в силу (4) минимизации энергии разбиения.

Не ограничивая общности, будем предполагать, что множество X упорядочено таким образом, что

- (а) $\langle x_1 - x_2 \rangle \geq \langle x_i - x_j \rangle$ для всех i и j ;
- (б) $\Phi(X_j, x_{j+1}) \geq \Phi(X_j, x_i)$ для всех $i \geq 2$ и $i > j$, где $X_j = \{x_1, \dots, x_j\}$;

величину $\Phi(X_j, x_{j+1})$ обозначим φ_{j+1} .

Приближенный алгоритм $\mathcal{A}_1$, начиная с разбиения $\Sigma_q(X_q) = \{\{x_1\}, \dots, \{x_q\}\}$, строит разбиение $\Sigma_q(X_{q+i})$, $i = 1, \dots, N - q$, применяя решающее правило $\mathcal{R}_1$ к точке x_{q+i} и разбиению $\Sigma_q(X_{q+i-1})$. В случае, когда для двух или более классов из $\Sigma_q(X_{q+i-1})$ достигается минимум потенциала $\Phi(K, x_{q+i})$, где $K \in \Sigma_q(X_{q+i-1})$, точка x_{q+i} относится к классу K с меньшим индексом.

Алгоритм $\mathcal{A}_1$ является градиентным (жадным) алгоритмом. К энергии разбиения $\Sigma_q(X_j)$ при классификации точки x_{j+1} добавляется минимально возможная часть от величины φ_{j+1} . Соответственно, максимально возможная часть от величины φ_{j+1} добавляется к $R(\Sigma_q(X_j))$. Более точно, при $j \geq q$ справедливо неравенство

$$R(\Sigma_q(X_{j+1})) \geq R(\Sigma_q(X_j)) + \frac{q-1}{q} \varphi_{j+1}. \quad (10)$$

В то же время для оптимальных разбиений, очевидно, выполняется неравенство

$$R(\bar{\Sigma}_q(X_{j+1})) \leq R(\bar{\Sigma}_q(X_j)) + \varphi_{j+1}. \quad (11)$$

Суммируя обе части неравенств (10) и (11) при $j = q, \dots, N-1$, получим, что

$$R(\Sigma_q(X_N)) \geq R(\Sigma_q(X_q)) + \frac{q-1}{q}(\varphi_{q+1} + \dots + \varphi_N), \quad (12)$$

$$R(\bar{\Sigma}_q(X_N)) \leq R(\bar{\Sigma}_q(X_q)) + (\varphi_{q+1} + \dots + \varphi_N). \quad (13)$$

Для произвольного $s \geq 2$

$$\sum_{j=2}^s \varphi_j = \sum_{j=2}^s \Phi(X_{j-1}, x_j) = \sum_{j=2}^s \sum_{i=1}^{j-1} \langle x_i - x_j \rangle = P(X_s). \quad (14)$$

Обозначим $R(\Sigma_q(X_N))$ через R_1 и $R(\bar{\Sigma}_q(X_N))$ через $\bar{R}$. Так как

$$\Sigma_q(X_q) = \bar{\Sigma}_q(X_q) = \{\{x_1\}, \dots, \{x_q\}\}$$

есть тривиальное разбиение, справедливо равенство

$$R(\Sigma_q(X_q)) = R(\bar{\Sigma}_q(X_q)) = P(X_q) = \varphi_2 + \dots + \varphi_q.$$

С учетом этого равенства, из (12) и (13) получаем следующую оценку отклонения $\delta_{A_1} = R_1/\bar{R}$ приближенного алгоритма A_1 :

$$\begin{aligned} R_1 = R(\Sigma_q(X_N)) &\geq R(\Sigma_q(X_q)) + \frac{q-1}{q}(\varphi_{q+1} + \dots + \varphi_N) \\ &> \frac{q-1}{q}(\varphi_2 + \dots + \varphi_N) \geq \frac{q-1}{q}R\bar{\Sigma}_q(X_N) = \frac{q-1}{q}\bar{R}, \end{aligned}$$

следовательно, $\delta_{A_1} > (q-1)/q$.

Пусть для некоторого $\alpha < 1$ выполняется неравенство $P(X_q) \geq \alpha P$, где P — потенциальная энергия всей системы точек X . Отметим, что для данного рационального α соответствующая целочисленная задача распознавания существования подмножества X_q , для которого $P(X_q) \geq \alpha P$, является NP-полной. Вычитая из (13) неравенство (12) и принимая во внимание (14), получаем следующую оценку абсолютной погрешности $\Delta_{A_1} = \bar{R} - R_1$ приближенного алгоритма A_1 :

$$\Delta_{A_1} = \bar{R} - R_1 \leq \frac{1}{q}(\varphi_{q+1} + \dots + \varphi_N) = \frac{1}{q}(P - P(X_q)) \leq \frac{1-\alpha}{q}P.$$

Таким образом, получено следующее утверждение.

Предложение 1. Для градиентного алгоритма кластеризации A_1 отношение $\delta_{A_1} = R_1/\bar{R}$ и абсолютная погрешность $\Delta_{A_1} = \bar{R} - R_1$ удовлетворяют неравенствам

$$\delta_{A_1} > \frac{q-1}{q}, \quad \Delta_{A_1} \leq \frac{1-\alpha}{q}P,$$

где P — потенциальная энергия всей системы X , q — число кластеров, $P(X_q)$ — энергия начального разбиения и $\alpha = P(X_q)/P$.

Заметим, что разбиение $\Sigma_q(X)$, которое строит алгоритм A_1 , может и не быть слабо $\mathcal{R}_1$ -разделимым. Тогда для получения слабо $\mathcal{R}_1$ -разделимого разбиения применяется процедура обмена векторов между классами, описанная выше, что приводит к увеличению отношения $R(\Sigma_q(X))/R(\bar{\Sigma}_q(X))$. Отметим, что для любого слабо $\mathcal{R}_1$ -разделимого разбиения $\Sigma_q(X)$ это отношение, очевидно, не меньше, чем $(q-1)/q$, что является необходимым, но не достаточным условием слабой $\mathcal{R}_1$ -разделимости.

Алгоритм A_1 прежде, чем начать склеивать точки в кластеры, выбирает прототипы центров будущих кластеров, множество X_q . Опишем другой градиентный алгоритм A_2 , который локально оптимальным образом склеивает кластеры, начиная с тривиального разбиения множества X на N кластеров. Так как при слиянии двух кластеров K_i и K_j в один энергия разбиения увеличивается на величину $\rho_{i,j} = P(K_i, K_j)$, на каждом шаге алгоритм A_2 объединяет в один кластер $K_l = K_i \cup K_j$ (l — новый индекс) кластеры K_i и K_j с минимальным значением $\rho_{i,j}$, новые величины $\rho_{k,l}$ для $k \neq i, j$ вычисляются по формуле $\rho_{k,l} = \rho_{k,i} + \rho_{k,j}$. После выполнения $N - q$ таких шагов алгоритм A_2 построит искомое разбиение множества X на q кластеров.

3. Кластеризация с минимальным потенциалом

В этом и следующем параграфах предполагаем, что метрический потенциал Φ дополнительно к аксиомам (1) и (2) удовлетворяет аксиоме суперпозиции:

- (3) для любого множества точек K величина $\Phi(K, y)$ достигает минимального значения при $y = \bar{x}$, где $\bar{x}$ — некоторая точка, называемая центром множества K , и для любого y выполняется равенство

$$\Phi(K, y) = \Phi(K, \bar{x}) + m\langle y - \bar{x} \rangle,$$

где $m = |K|$.

Из аксиомы суперпозиции непосредственно вытекает, что центр $\bar{x}$ множества K определяется однозначно.

Аксиоме суперпозиции удовлетворяют, например, следующие метрические потенциалы:

$\langle x - y \rangle = \rho_E^2(x - y)$, где $x, y \in R^n$, а ρ_E — евклидова метрика (этот случай наиболее интенсивно изучается в работах по кластеризации);

$\langle x - y \rangle = (x - y, G(x - y))$, где $(\cdot, \cdot)$ — скалярное произведение векторов из R^n , а G — симметрическая положительно определенная матрица размера $n \times n$;

$\langle x - y \rangle = \rho_H(x - y)$, где $x, y \in E_2^n = \{0, 1\}^n$, а ρ_H — метрика Хемминга.

Пусть, как и раньше, Σ_q — разбиение множества X на q кластеров $K_1, \dots, K_q$, σ_q — соответствующее разбиение множества индексов на множества $k_1, \dots, k_q$ и $m_j = |K_j|$, $j = 1, \dots, q$.

Для кластера K_s с центром $\bar{x}_s$ величину

$$\Phi_s = \Phi(K_s, \bar{x}_s) = \sum_{i \in k_s} \langle x_i - \bar{x}_s \rangle \quad (15)$$

называем потенциалом кластера. Аксиома суперпозиции принимает вид

$$\Phi(K_s, y) = \Phi_s + m_s \langle y - \bar{x}_s \rangle. \quad (16)$$

Потенциал Φ_s кластера K_s связан с его потенциальной энергией P_s , определенной в предыдущем параграфе, равенством $P_s = m_s \Phi_s$. Действительно,

$$\begin{aligned} P_s &= \frac{1}{2} \sum_{i \in k_s} \sum_{j \in k_s} \langle x_i - x_j \rangle = \frac{1}{2} \sum_{i \in k_s} \Phi(K_s, x_i) \\ &= \frac{1}{2} \sum_{i \in k_s} (\Phi_s + m_s \langle x_i - \bar{x}_s \rangle) = \frac{1}{2} m_s \Phi_s + \frac{1}{2} m_s \sum_{i \in k_s} \Phi(K_s, \bar{x}_s) \\ &= \frac{1}{2} m_s \Phi_s + \frac{1}{2} m_s \Phi_s = m_s \Phi_s. \end{aligned} \quad (17)$$

Аналогичным образом можно преобразовать выражение для потенциальной энергии двух кластеров K_r и K_s относительно друг друга:

$$\begin{aligned} P(K_r, K_s) &= \sum_{i \in k_r} \sum_{j \in k_s} \langle x_i - x_j \rangle = \sum_{i \in k_r} \Phi(K_s, x_i) \\ &= \sum_{i \in k_r} (\Phi_s + m_s \langle x_i - \bar{x}_s \rangle) = m_r \Phi_s + m_s \Phi(K_r, \bar{x}_s) \\ &= m_r \Phi_s + m_s \Phi_r + m_r m_s \langle \bar{x}_r - \bar{x}_s \rangle. \end{aligned} \quad (18)$$

Вычислим потенциал Φ всей системы точек X :

$$\begin{aligned} \Phi &= \Phi(X, \bar{x}) = \Phi(K_1 \cup \dots \cup K_q, \bar{x}) \\ &= \sum_{i=1}^q \Phi(K_i, \bar{x}) = \sum_{i=1}^q (\Phi_i + m_i \langle \bar{x}_i - \bar{x} \rangle) = \sum_{i=1}^q \Phi_i + \sum_{i=1}^q m_i \langle \bar{x}_i - \bar{x} \rangle, \end{aligned} \quad (19)$$

где $\bar{x}$ — центр множества X , $\bar{x}_i$ — центр кластера K_i , $i = 1, \dots, q$.

Потенциал Φ можно вычислить и по-другому, исходя из потенциальной энергии P всей системы. Используя (4) и (18), получим, что

$$\begin{aligned} \Phi &= \frac{1}{N} P = \frac{1}{2N} \sum_{i=1}^q \sum_{j=1}^q P(K_i, K_j) = \frac{1}{2N} \sum_{i=1}^q \sum_{j=1}^q (m_i \Phi_j + m_j \Phi_i + m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle) \\ &= \frac{1}{2N} \left(\sum_{i=1}^q m_i \sum_{j=1}^q \Phi_j + \sum_{j=1}^q m_j \sum_{i=1}^q \Phi_i + \sum_{i=1}^q \sum_{j=1}^q m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle \right) \\ &= \sum_{i=1}^q \Phi_i + \frac{1}{N} \sum_{i < j} m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle. \end{aligned} \quad (20)$$

Тем самым установлено тождество для центров

$$\frac{1}{N} \sum_{i < j} m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle = \sum_{i=1}^q m_i \langle \bar{x}_i - \bar{x} \rangle. \quad (21)$$

Слагаемое $\sum_{i=1}^q \Phi_i$ в правой части равенства (20) характеризует компактность кластеров в разбиении Σ_q , в то время как слагаемое $\frac{1}{N} \sum_{i < j} m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle$ характеризует степень различия между кластерами. Потенциал Φ не зависит от разбиения Σ_q , поэтому минимизируя величину $\Phi(\Sigma_q) = \sum_{i=1}^q \Phi_i$, мы одновременно максимизируем величину $\sum_{i < j} m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle$.

Величину $\Phi(\Sigma_q)$ будем называть потенциалом разбиения Σ_q , а разбиение $\bar{\Sigma}_q$ с минимальным значением потенциала — оптимальным (относительно потенциала). Точно так же, как и в случае с потенциальной энергией, показывается, что функция $\bar{\Phi}(\Sigma_q)$ строго убывающая.

Найдем потенциал кластера K_l , образованного слиянием или разбиением двух других кластеров K_i и K_j . Пусть $K_l = K_i \cup K_j$. Тогда из (5), используя (17) и (18), получим, что

$$\begin{aligned} \Phi_l &= \frac{1}{m_l} P_l = \frac{1}{m_i + m_j} (P_i + P_j + P(K_i, K_j)) \\ &= \frac{1}{m_i + m_j} (m_i \Phi_i + m_j \Phi_j + m_j \Phi_i + m_i \Phi_j + m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle) \\ &= \Phi_i + \Phi_j + \frac{m_i m_j}{m_i + m_j} \langle \bar{x}_i - \bar{x}_j \rangle. \end{aligned} \quad (22)$$

В частности, если $K_i = \{x\}$ — одноточечный кластер, то

$$\Phi'_j = \Phi_j + \frac{m_j}{m_j + 1} \langle x - \bar{x}_j \rangle, \quad (23)$$

где Φ'_j означает потенциал кластера $K_j \cup \{x\}$.

Аналогично, если $K_l = K_i \setminus K_j$ и $K_j \subset K_i$, то

$$\Phi_l = \Phi_i - \Phi_j - \frac{m_i m_j}{m_i - m_j} \langle \bar{x}_i - \bar{x}_j \rangle, \quad (24)$$

и в случае одноточечного кластера $K_j = \{x\}$ при $m_i \geq 2$

$$\Phi'_i = \Phi_i - \frac{m_i}{m_i - 1} \langle x - \bar{x}_i \rangle, \quad (25)$$

где Φ'_i — потенциал кластера $K_i \setminus \{x\}$.

С рассматриваемым потенциальным критерием кластеризации естественным образом ассоциируется следующее решающее правило:

$$\mathcal{R}_2: x \in K_i \iff \Phi(\bar{x}_i, x) < \Phi(\bar{x}_j, x) \iff \langle \bar{x}_i - x \rangle < \langle \bar{x}_j - x \rangle, \quad j \neq i.$$

Точка x , подлежащая классификации, относится к тому из кластеров (классов), относительно центра которого она имеет меньший потенциал. В случае, когда минимум $\Phi(\bar{x}_i, x)$ достигается на нескольких центрах, происходит отказ от классификации.

Теорема 2. *Оптимальные разбиения $\bar{\Sigma}_q$ являются $\mathcal{R}_2$ -разделимыми.*

Доказательство. Пусть $\bar{\Sigma}_q$ — оптимальное разбиение. Очевидно, что решающее правило $\mathcal{R}_2$ правильно классифицирует точки из одноэлементных классов. Пусть для кластера K_i выполняется неравенство $m_i \geq 2$ и пусть x — произвольная точка из K_i . Выберем произвольный кластер K_j , $j \neq i$, и образуем новое разбиение Σ_q на q кластеров, полученное из $\bar{\Sigma}_q$ в результате переноса точки x из кластера K_i в кластер K_j , т.е.

$$\Sigma_q = \{K_1, \dots, K_i \setminus \{x\}, \dots, K_j \cup \{x\}, \dots, K_q\}.$$

В силу (22) и (24) изменение $\Delta\Phi$ потенциала разбиения равно

$$\begin{aligned} \Delta\Phi &= \Phi(\Sigma_q) - \Phi(\bar{\Sigma}_q) = \Phi'_i - \Phi_i + \Phi'_j - \Phi_j \\ &= -\frac{m_i}{m_i - 1} \langle \bar{x}_i - x \rangle + \frac{m_j}{m_j + 1} \langle \bar{x}_j - x \rangle \geq 0. \end{aligned} \quad (26)$$

Так как $\bar{\Sigma}_q$ — оптимальное разбиение, то $\Delta\Phi \geq 0$; принимая также во внимание, что $m_i/(m_i - 1) > 1$, а $m_j/(m_j + 1) < 1$, из (26) окончательно получим, что

$$\langle \bar{x}_i - x \rangle < \langle \bar{x}_j - x \rangle,$$

что означает правильную классификацию точки x решающим правилом $\mathcal{R}_2$. Теорема доказана.

Таким образом, соотношение (26) является необходимым условием оптимальности разбиения. В случае евклидовой полуметрики $\langle x - y \rangle = \rho_E^2(x - y)$ это условие хорошо известно и используется в итерационных процедурах кластеризации.

В качестве начального приближения в таких итерационных процедурах можно использовать разбиения, которые строят градиентные алгоритмы. Рассмотрим два таких алгоритма $\mathcal{B}_1$ и $\mathcal{B}_2$.

Алгоритм $\mathcal{B}_1$ определяет разбиение $\Sigma_q(X_q) = \{\{x_1\}, \dots, \{x_q\}\}$ точно так же, как и алгоритм $\mathcal{A}_1$ из предыдущего параграфа, и далее строит разбиение $\Sigma_q(X_{q+i})$, $i = 1, \dots, N - q$, применяя решающее правило $\mathcal{R}_2$ к точке x_{q+i} и разбиению $\Sigma_q(X_{q+i-1})$.

Обозначим $\bar{x}_i^j$, $i = 1, \dots, q$, $j = q, \dots, N - 1$, центры кластеров в разбиении $\Sigma_q(X_j)$, так что, например, $\bar{x}_i^q = x_i$, $i = 1, \dots, q$, и пусть

$$\psi_{q+1} = \sum_{i=1}^q \langle x_{j+1} - \bar{x}_i^j \rangle.$$

Учитывая (23), для изменения потенциала разбиений получим оценку

$$\Phi(\Sigma_q(X_{j+1})) - \Phi(\Sigma_q(X_j)) = \frac{m_s^j}{m_s^j + 1} \langle x_{j+1} - \bar{x}_s^j \rangle < \langle x_{j+1} - \bar{x}_s^j \rangle \leq \frac{\psi_{j+1}}{q}, \quad (27)$$

где центр $\bar{x}_s^j$ кластера K_s^j из разбиения $\Sigma_q(X_j)$ — ближайший центр к точке x_{j+1} . Так как $\Phi(\Sigma_q(X_q)) = 0$, отсюда следует, что для потенциала окончательного разбиения справедлива оценка

$$\Phi(\Sigma_q(X)) < \frac{1}{q} \sum_{j=q}^{N-1} \psi_{j+1}. \quad (28)$$

Для оценки правой части неравенства (28) нам понадобится следующее рассуждение. Пусть K — произвольный кластер с центром в точке $\bar{x}$. При добавлении к кластеру некоторой точки y его новый центр обозначим $\bar{x}'$. Для произвольной точки z

$$\Phi(K \cup \{y\}, z) = \Phi(K, z) + \langle y - z \rangle = \Phi + m\langle z - \bar{x} \rangle + \langle y - z \rangle, \quad (29)$$

где, как обычно, Φ — потенциал кластера K и $m = |K|$. С другой стороны, обозначая Φ' потенциал кластера $K \cup \{y\}$ и используя (22), получим, что

$$\Phi(K \cup \{y\}, z) = \Phi' + (m+1)\langle z - \bar{x}' \rangle = \Phi + \frac{m}{m+1}\langle y - \bar{x} \rangle + (m+1)\langle z - \bar{x}' \rangle. \quad (30)$$

Сравнивая (29) и (30), приходим к неравенству

$$\begin{aligned} \langle z - \bar{x}' \rangle &= \frac{m}{m+1}\langle z - \bar{x} \rangle + \frac{1}{m+1}\langle z - y \rangle - \frac{m}{(m+1)^2}\langle y - \bar{x} \rangle \\ &< \langle z - \bar{x} \rangle + \frac{1}{2}\langle z - y \rangle. \end{aligned} \quad (31)$$

Применяя (31) к (28), получим, что

$$\begin{aligned} \Phi(\Sigma_q) &< \frac{1}{q} \sum_{j=q}^{N-1} \psi_{j+1} = \frac{1}{q} \sum_{j=q}^{N-1} \sum_{i=1}^q \langle x_{j+1} - \bar{x}_i^j \rangle \\ &< \frac{1}{q} \left(\sum_{j=q}^{N-1} \sum_{i=1}^q \langle x_{j+1} - x_i \rangle + \frac{1}{2} \sum_{j=q}^{N-1} \sum_{k=q+1}^j \langle x_{j+1} - x_k \rangle \right) \\ &= \frac{1}{q} \left(\sum_{j=q}^{N-1} \Phi(X_q, x_{j+1}) + \frac{1}{2} P(X \setminus X_q) \right) = \frac{1}{q} \left(P(X_q, X \setminus X_q) + \frac{1}{2} P(X \setminus X_q) \right) \\ &= \frac{1}{q} \left(P - P(X_q) - \frac{1}{2} P(X \setminus X_q) \right), \end{aligned} \quad (32)$$

где $P = P(X)$ — энергия всего множества точек.

Отметим, что последнее неравенство в (27) довольно грубое, так что (32) может вырождаться в тривиальную оценку потенциала разбиения, которое строит алгоритм $\mathcal{B}_1$. Можно показать, что, по крайней мере, при $q/N > 1 - \alpha$, где $\alpha = P(X_q)/P$, она будет нетривиальной.

Предложение 2. При $q/N > 1 - \alpha$, где $\alpha = P(X_q)/P$, $P = P(X)$ — энергия множества X , а $P(X_q)$ — энергия начального разбиения, потенциал $\Phi(\Sigma_q)$ разбиения Σ_q , которое строит алгоритм кластеризации $\mathcal{B}_1$, удовлетворяет неравенству

$$\Phi(\Sigma_q) < \frac{1}{q} \left(P - P(X_q) - \frac{1}{2} P(X \setminus X_q) \right).$$

Перейдем к описанию алгоритма $\mathcal{B}_2$. Этот алгоритм опирается на формулу (22). Пусть K_i и K_j — два кластера. При их слиянии потенциал разбиения увеличивается на величину $m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle / (m_i + m_j)$. Положим

$$\rho_{i,j} = \frac{m_i m_j}{m_i + m_j} \langle \bar{x}_i - \bar{x}_j \rangle.$$

Алгоритм $\mathcal{B}_2$, отправляясь от тривиального разбиения множества X на N классов, на каждом шаге сливает два кластера K_i и K_j с минимальным значением $\rho_{i,j}$ в один новый кластер $K_l = K_i \cup K_j$ (l — новый индекс). При этом для $k \neq i, j$ вычисляются новые величины расстояний

$$\rho_{k,l} = \frac{(m_i + m_k)\rho_{k,i} + (m_j + m_k)\rho_{k,j} - m_k\rho_{i,j}}{m_i + m_j + m_k}.$$

После выполнения $N - q$ таких шагов, алгоритм $\mathcal{B}_2$ построит разбиение множества X на q кластеров.

4. Кластеризация с минимальным размером

Для произвольного кластера K_s из разбиения Σ_q величину

$$w_s = \frac{\Phi_s}{m_s} = \frac{1}{m_s} \sum_{i \in K_s} \langle x_i - \bar{x}_s \rangle, \quad (33)$$

где Φ_s и $\bar{x}_s$ — потенциал и центр кластера K_s , а $m_s = |K_s|$, будем называть размером кластера.

С каждым кластером можно связать два множества, ядро и оболочку кластера. Ядро $\varepsilon(K_s)$, порождаемое кластером K_s , есть часть пространства, примыкающая к его центру:

$$\varepsilon(K_s) = \{x \mid \langle x - \bar{x}_s \rangle \leq w_s\},$$

в то время как оболочка кластера, есть та его часть, которая лежит вне внутренности ядра:

$$\hat{K}_s = K_s \setminus \varepsilon_0(K_s),$$

где внутренность ядра — это множество

$$\varepsilon_0(K_s) = \{x \mid \langle x - \bar{x}_s \rangle < w_s\}.$$

Отметим, что если внутренность ядра кластера может быть и пустой (например, для одноточечных кластеров), то его оболочка $\hat{K}_s$ всегда непуста. Действительно, пусть это не так и каждая точка x из кластера K_s расположена во внутренности ядра. Это означает, что $\langle x - \bar{x}_s \rangle < w_s$ для всех $x \in K_s$. Суммируя, получаем противоречие:

$$\Phi_s = \sum_{i \in K_s} \langle x_i - \bar{x}_s \rangle < m_s w_s = \Phi_s.$$

Аналогично получается, что $K_s \cap \varepsilon(K_s) \neq \emptyset$.

Если Σ_q — разбиение, то его размером $w(\Sigma_q)$ будем называть величину

$$w(\Sigma_q) = \sum_{s=1}^q w_s.$$

Разбиение $\bar{\Sigma}_q$ с минимальным значением размера называется оптимальным (относительно размера).

Из определения размера кластера и из формул (22)–(25) непосредственно получаются выражения для размера кластера, образованного слиянием или разбиением двух других кластеров, в том числе когда один из кластеров одноточечный.

Сохраняя обозначения параграфа 3, получаем, что

$$w_l = \frac{m_i w_i + m_j w_j}{m_i + m_j} + \frac{m_i m_j}{(m_i + m_j)^2} \langle \bar{x}_i - \bar{x}_j \rangle, \quad (34)$$

$$w'_j = \frac{m_j w_j}{m_j + 1} + \frac{m_j}{(m_j + 1)^2} \langle \bar{x}_j - x \rangle, \quad (35)$$

$$w_l = \frac{m_i w_i - m_j w_j}{m_i - m_j} - \frac{m_i m_j}{(m_i - m_j)^2} \langle \bar{x}_i - \bar{x}_j \rangle, \quad (36)$$

$$w'_i = \frac{m_i w_i}{m_i - 1} - \frac{m_i}{(m_i - 1)^2} \langle \bar{x}_i - x \rangle. \quad (37)$$

Несмотря на то, что для размера $w(X)$ всей системы точек нет прямых аналогов формул (4) и (19) из предыдущих параграфов, свойство монотонности функции $w(\bar{\Sigma}_q)$ как функции от q выполняется и проверяется по той же схеме.

Действительно, при $q = N$ справедливо равенство $w(\bar{\Sigma}_N) = 0$. При $q \leq N - 1$ найдем в разбиении $\bar{\Sigma}_q = \{K_1, \dots, K_q\}$ неодноточечный кластер K_i и в нем точку x из оболочки кластера K_i . Образует разбиение на $q + 1$ кластер

$$\Sigma_{q+1} = \{K_1, \dots, K_i \setminus \{x\}, \dots, K_q, \{x\}\}.$$

Ясно, что

$$\begin{aligned} w(\bar{\Sigma}_q) - w(\Sigma_{q+1}) &= w_i - w'_i = \frac{1}{m_i - 1} \left(\frac{m_i}{m_i - 1} \langle \bar{x}_i - x \rangle - w_i \right) \\ &> \frac{1}{m_i - 1} (\langle \bar{x}_i - x \rangle - w_i) \geq 0, \end{aligned}$$

так как $x \in \hat{K}_i$. Отсюда и из неравенства $w(\Sigma_{q+1}) \geq w(\bar{\Sigma}_{q+1})$ окончательно находим, что $w(\bar{\Sigma}_q) > w(\bar{\Sigma}_{q+1})$.

Нам понадобится следующее свойство ядер кластеров в оптимальном разбиении $\bar{\Sigma}_q$.

Лемма 1. *Ядра разных кластеров в оптимальном разбиении не пересекаются.*

Доказательство. Пусть это не так и для кластеров K_i и K_j , $i \neq j$, из оптимального разбиения $\bar{\Sigma}_q = \{K_1, \dots, K_q\}$ найдется $x \in \varepsilon(K_i) \cap \varepsilon(K_j)$. В частности, это означает, что

$$\langle x - \bar{x}_i \rangle \leq w_i, \quad \langle x - \bar{x}_j \rangle \leq w_j. \quad (38)$$

В разбиении $\bar{\Sigma}_q$ сольем два кластера K_i и K_j в один, $K_l = K_i \cup K_j$, и рассмотрим получившиеся разбиение Σ_{q-1} . В силу монотонности

$$w(\Sigma_{q-1}) \geq w(\bar{\Sigma}_{q-1}) > w(\bar{\Sigma}_q). \quad (39)$$

С другой стороны, используя (34), получим, что

$$\begin{aligned} w(\Sigma_{q-1}) - w(\bar{\Sigma}_q) &= w_l - w_i - w_j = \frac{m_i w_i + m_j w_j}{m_i + m_j} + \frac{m_i w_j}{(m_i + m_j)^2} \langle \bar{x}_i - \bar{x}_j \rangle - w_i - w_j \\ &= \frac{1}{m_i + m_j} \left(\frac{m_i w_j}{m_i + m_j} \langle \bar{x}_i - \bar{x}_j \rangle - m_j w_i - m_i w_j \right). \end{aligned} \quad (40)$$

Слагаемое $m_i m_j \langle \bar{x}_i - \bar{x}_j \rangle / (m_i + m_j)$ есть, по определению, потенциал системы точек K' , состоящей из m_j точек x_i и m_i точек x_j , центр которой обозначим $\bar{x}'$. Отсюда (отметим, что x может не быть центром кластера K') следует, что

$$\frac{m_i m_j}{m_i + m_j} \langle \bar{x}_i - \bar{x}_j \rangle = \Phi(K', \bar{x}') \leq \Phi(K', x) = m_j \langle \bar{x}_i - x \rangle + m_i \langle \bar{x}_j - x \rangle.$$

Отсюда и из (38), подставляя полученную оценку в (40), находим, что выполняется неравенство $w(\Sigma_{q-1}) - w(\bar{\Sigma}_q) \leq 0$, что противоречит (39). Лемма доказана.

Лемма дает необходимое условие оптимальности: точки из ядра кластера должны в случае оптимального разбиения заведомо классифицироваться как принадлежащие рассматриваемому кластеру. Так как точки кластера из его оболочки характеризуются неравенством $\langle x - \bar{x} \rangle \geq w$, одним из подходящих решающих правил для этого случая (с точностью до сомножителя, учитывающего мощность кластера) будет правило

$$\mathcal{R}_3: x \in K_i \iff \frac{1}{m_i} (\langle x - \bar{x}_i \rangle - w_i) < \frac{1}{m_j} (\langle x - \bar{x}_j \rangle - w_j), \quad j \neq i.$$

Теорема 3. *Оптимальные разбиения $\bar{\Sigma}_q$ являются $\mathcal{R}_3$ -разделимыми.*

Доказательство. Пусть $\bar{\Sigma}_q$ — оптимальное разбиение. Как следует из леммы, решающее правило $\mathcal{R}_3$ правильно классифицирует точки из ядер кластеров, в частности, и точки из одноэлементных кластеров. Рассмотрим теперь произвольную точку x из оболочки некоторого кластера K_i , где $m_i \geq 2$. Выберем произвольный кластер K_j , $j \neq i$, и образуем новое разбиение Σ_q , полученное из разбиения $\bar{\Sigma}_q$ в результате переноса точки x из кластера K_i в кластер K_j . Используя (35) и (37), вычислим изменение Δw размера разбиения:

$$\begin{aligned} \Delta w &= w(\Sigma_q) - w(\bar{\Sigma}_q) = w'_j - w_j + w'_i - w_i \\ &= \frac{m_j}{m_j + 1} w_j - w_j + \frac{m_j}{(m_j + 1)^2} \langle \bar{x}_j - x \rangle + \frac{m_i}{m_i - 1} w_i - w_i + \frac{m_i}{(m_i - 1)^2} \langle \bar{x}_i - x \rangle \\ &= \frac{1}{m_j + 1} \left(\frac{m_j}{m_j + 1} \langle \bar{x}_j - x \rangle - w_j \right) - \frac{1}{m_i - 1} \left(\frac{m_i}{m_i - 1} \langle \bar{x}_i - x \rangle - w_i \right) \geq 0. \end{aligned}$$

Так как $\bar{\Sigma}_q$ — оптимальное разбиение, то $\Delta w \geq 0$.

Учитывая, что $m_j / (m_j + 1) < 1$ и $m_i / (m_i - 1) > 1$, и что $\langle \bar{x}_i - x \rangle - w_i \geq 0$, получаем оценку

$$\Delta w < \frac{1}{m_j} (\langle \bar{x}_j - x \rangle - w_j) - \frac{1}{m_i} (\langle \bar{x}_i - x \rangle - w_i).$$

Следовательно,

$$\frac{1}{m_i}(\langle x - \bar{x}_i \rangle - w_i) < \frac{1}{m_j}(\langle x - \bar{x}_j \rangle - w_j),$$

что означает правильную классификацию точки x решающим правилом $\mathcal{R}_3$. Теорема доказана.

Пусть $\Sigma_q = \{K_1, \dots, K_q\}$ — произвольное разбиение множества X на q кластеров. Отправляясь от разбиения Σ_q , можно строить разбиение Σ_q^π , в котором ядра разных кластеров не пересекаются, используя следующую процедуру π .

Для произвольных кластеров K_i и K_j из Σ_q , $i \neq j$, определим $\rho_{i,j} = w_l - w_i - w_j$, где w_l — размер кластера $K_l = K_i \cup K_j$, образованного слиянием рассматриваемых кластеров K_i и K_j . Находим в разбиении два кластера с минимальным значением $\rho_{i,j}$. Не ограничивая общности, считаем, что для кластеров K_i и K_j величина $\rho_{i,j}$ минимальна. В случае $\rho_{i,j} \leq 0$ строим новое разбиение Σ_q' : сливаем кластеры K_i и K_j в один кластер K_l и далее точку x из оболочки кластера K_l с максимальным значением $\langle x - \bar{x}_l \rangle$ выделяем в отдельный одноточечный кластер. Таким образом, процедура π в случае $\rho_{i,j} \leq 0$ на каждом шаге заменяет в разбиении два кластера K_i и K_j на два новых, K_i' и $K_j' = \{x\}$, где $K_i' = K_l \setminus \{x\} = (K_i \cup K_j) \setminus \{x\}$, после чего π применяется к новому разбиению. Если после выполнения k таких шагов приходим к разбиению Σ_q^k , для любых двух кластеров K_i^k и K_j^k которого $\rho_{i,j} > 0$, то выполнение процедуры заканчивается и $\Sigma_q^\pi = \Sigma_q^k$. Ясно, что в разбиении Σ_q^π ядра разных кластеров не пересекаются. Действительно, в противном случае для кластеров K_i^k и K_j^k из Σ_q^k , для которых $\varepsilon(K_i) \cup \varepsilon(K_j) \neq \emptyset$, с использованием (38) и (40), получаем, что $\rho_{i,j} \leq 0$, что противоречит условию остановки процедуры π .

Покажем теперь, что процедура π останавливается за конечное число шагов. Для этого оценим изменение $\Delta = w(\Sigma_q) - w(\Sigma_q')$ размера разбиения при выполнении одного шага процедуры. Используя (37) и описание π , получим, что

$$\begin{aligned} \Delta &= w(\Sigma_q) - w(\Sigma_q') = w_i + w_j - w_i' = w_i + w_j - \frac{m_l w_l}{m_l - 1} + \frac{m_l}{(m_l - 1)^2} \langle x - \bar{x}_l \rangle \\ &= -\rho_{i,j} + \frac{1}{m_l - 1} \left(\frac{m_l}{m_l - 1} \langle x - \bar{x}_l \rangle - w_l \right) \\ &\geq -\rho_{i,j} + \frac{1}{m_l - 1} \left(\frac{\langle x - \bar{x}_l \rangle}{m_l - 1} + \langle x - \bar{x}_l \rangle - w_l \right) \\ &\geq \frac{\langle x - \bar{x}_l \rangle}{(m_l - 1)^2} \geq \frac{w_l}{(m_l - 1)^2} = \frac{1}{m_l^2 (m_l - 1)^2} \sum_{r,s \in K_s, r < s} \langle x_r - x_s \rangle \\ &\geq \frac{m_l (m_l - 1) \delta}{2 m_l^2 (m_l - 1)^2} = \frac{\delta}{2 m_l (m_l - 1)} > \frac{\delta}{2 N^2}, \end{aligned}$$

где

$$\delta = \min_{i \neq j} \langle x_i - x_j \rangle, \quad x_i, x_j \in X, \quad N = |X|.$$

Таким образом, процедура π на каждом шаге уменьшает размер разбиения больше, чем на величину $\delta/(2N^2)$. Ясно, что число шагов заведомо не превосходит $2N^2 w(X)/\delta$, где $w(X)$ — размер всей системы точек.

Для построения разбиения Σ_q , близкого к оптимальному, как и ранее, организуется итерационная процедура обмена векторов, основанная на необходимом условии оптимальности.

В качестве приемлемых разбиений до применения итерационной процедуры обмена векторов можно использовать разбиения, которые строят градиентные алгоритмы. Рассмотрим составные градиентные алгоритмы кластеризации $\mathcal{C}_1$ и $\mathcal{C}_2$, включающие в себя описанную выше процедуру π .

В случае алгоритма $\mathcal{C}_1$ определяем разбиение $\Sigma_q(X_q)$ точно так же, как и для алгоритма $\mathcal{A}_1$. Далее, для $i = 1, \dots, N - q$, применяя решающее правило $\mathcal{R}_3$ к точке x_{q+i} и разбиению $\Sigma_q(X_{q+i-1})$, строим разбиение $\tilde{\Sigma}_q(X_{q+i})$. К разбиению $\tilde{\Sigma}_q(X_{q+i})$ применяем процедуру π и окончательно полагаем $\Sigma_q(X_{q+i}) = \tilde{\Sigma}_q^\pi(X_{q+i})$.

В случае алгоритма $\mathcal{C}_2$, отправляясь от тривиального разбиения $\tilde{\Sigma}_N$ множества X на N одноточечных кластеров, так же, как и в случае процедуры π , находим в разбиении $\tilde{\Sigma}_{N+1-i}$, $i = 1, \dots, N - q$, два кластера K_i и K_j с минимальным значением $\rho_{i,j} = w_l - w_i - w_j$, где w_l — размер кластера $K_l = K_i \cup K_j$, и объединяем их (l — новый индекс). Для полученного таким образом разбиения $\tilde{\Sigma}_{N-i}$ вычисляются новые значения $\rho_{k,l}$ для $k \neq i, j$. Непосредственное вычисление приводит к формулам

$$\rho_{k,l} = \frac{(m_i + m_k)\rho_{i,k} + (m_j + m_k)\rho_{j,k} - m_k\rho_{i,j}}{m_i + m_j + m_k} - \frac{m_i m_j m_k}{m_i + m_j + m_k} \left(\frac{\langle \bar{x}_i - \bar{x}_k \rangle}{m_i + m_k} + \frac{\langle \bar{x}_j - \bar{x}_k \rangle}{m_j + m_k} + \frac{\langle \bar{x}_i - \bar{x}_j \rangle}{m_i + m_j} \right).$$

После выполнения $N - q$ таких шагов приходим к разбиению $\tilde{\Sigma}_q$, после чего применяется процедура π . Окончательно полагаем $\Sigma_q = \tilde{\Sigma}_q^\pi$.

Оба алгоритма $\mathcal{C}_1$ и $\mathcal{C}_2$, таким образом, строят (каждый свое) разбиение Σ_q , в котором ядра разных кластеров не пересекаются. Отсюда извлекается оценка размера разбиения Σ_q . Действительно, для любых i и j , $i \neq j$, так как $\bar{x}_i \in \varepsilon(K_i)$ и $\bar{x}_j \notin \varepsilon(K_j)$, получим, что $\langle \bar{x}_i - \bar{x}_j \rangle > w_j$. Аналогично, $\langle \bar{x}_i - \bar{x}_j \rangle > w_i$. Складывая эти неравенства, находим, что

$$\langle \bar{x}_i - \bar{x}_j \rangle > \frac{1}{2}(w_i + w_j). \quad (41)$$

Суммирование неравенства (41) по всем $i \neq j$ приводит к оценкам

$$\begin{aligned} \sum_i \sum_j \langle \bar{x}_i - \bar{x}_j \rangle &> \frac{1}{2} \left(\sum_i \sum_j (w_i + w_j) - 2 \sum_i w_i \right), \\ 2P(\tilde{X}_q) &> \frac{1}{2} \left(\sum_i q w_i + \sum_i w(\Sigma_q) - 2w(\Sigma_q) \right) = (q - 1)w(\Sigma_q), \\ \frac{2q^2}{q - 1} w(\tilde{X}_q) &> w(\Sigma_q), \end{aligned}$$

где $\tilde{X}_q = \{\bar{x}_1, \dots, \bar{x}_q\}$ — множество центров кластеров из разбиения Σ_q .

Таким образом, доказано следующее утверждение.

Предложение 3. Размер $w(\Sigma_q)$ разбиений, которые строят алгоритмы кластеризации $\mathcal{C}_1$ и $\mathcal{C}_2$ удовлетворяет неравенству

$$w(\Sigma_q) < \frac{2q^2}{q-1} w(\tilde{X}_q),$$

где $\tilde{X}_q$ — соответствующее множество центров кластеров из разбиения.

Отметим, что в случае евклидовой полуметрики $\langle x - y \rangle = \rho_E^2(x - y)$, неравенство (41) может быть усилено до неравенства $\langle \bar{x}_i - \bar{x}_j \rangle > w_i + w_j$, что приводит к оценке

$$w(\Sigma_q) < \frac{q^2}{q-1} w(\tilde{X}_q).$$

Список литературы

1. Болотов А. А., Шайеб А., О метрических алгоритмах классификации. Докл. АН СССР (1987) **297**, №3, 527–530.

Статья поступила 04.10.96.