Лекция 11. Методы и алгоритмы анализа структуры многомерных данных. Кластерный анализ.

Кластерный анализ предназначен для разбиения множества объектов на заданное или неизвестное число классов на основании некоторого математического критерия качества классификации (cluster (англ.) — гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством).

Критерий качества кластеризации в той или иной мере отражает следующие неформальные требования:

а) внутри групп объекты должны быть тесно связаны между собой;
б) объекты разных групп должны быть далеки друг от друга;
в) при прочих равных условиях распределения объектов по группам должны быть равномерными.

Требования а) и б) выражают стандартную концепцию компактности классов разбиения; требование в) состоит в том, чтобы критерий не навязывал объединения отдельных групп объектов.

Узловым моментом в кластерном анализе считается выбор метрики (или меры близости объектов), от которого решающим образом зависит окончательный вариант разбиения объектов на группы при заданном алгоритме разбиения.

В каждой конкретной задаче этот выбор производится по-своему, с учетом главных целей исследования, физической и статистической природы используемой информации и т. п. При применении экстенсиональных методов распознавания, как было показано в предыдущих разделах, выбор метрики достигается с помощью специальных алгоритмов преобразования исходного пространства признаков.

Другой важной величиной в кластерном анализе является расстояние между целыми группами объектов.

Приведем примеры наиболее распространенных расстояний и мер близости, характеризующих взаимное расположение отдельных групп объектов.

Пусть w_i — *i*-я группа (класс, кластер) объектов, N_i — число объектов, образующих группу w_i , вектор μ_i — среднее арифметическое объектов, входящих в w_i (другими словами: μ_i — “центр тяжести” *i*-й группы), а $q(w_i, w_m)$ — расстояние между группами w_i и w_m

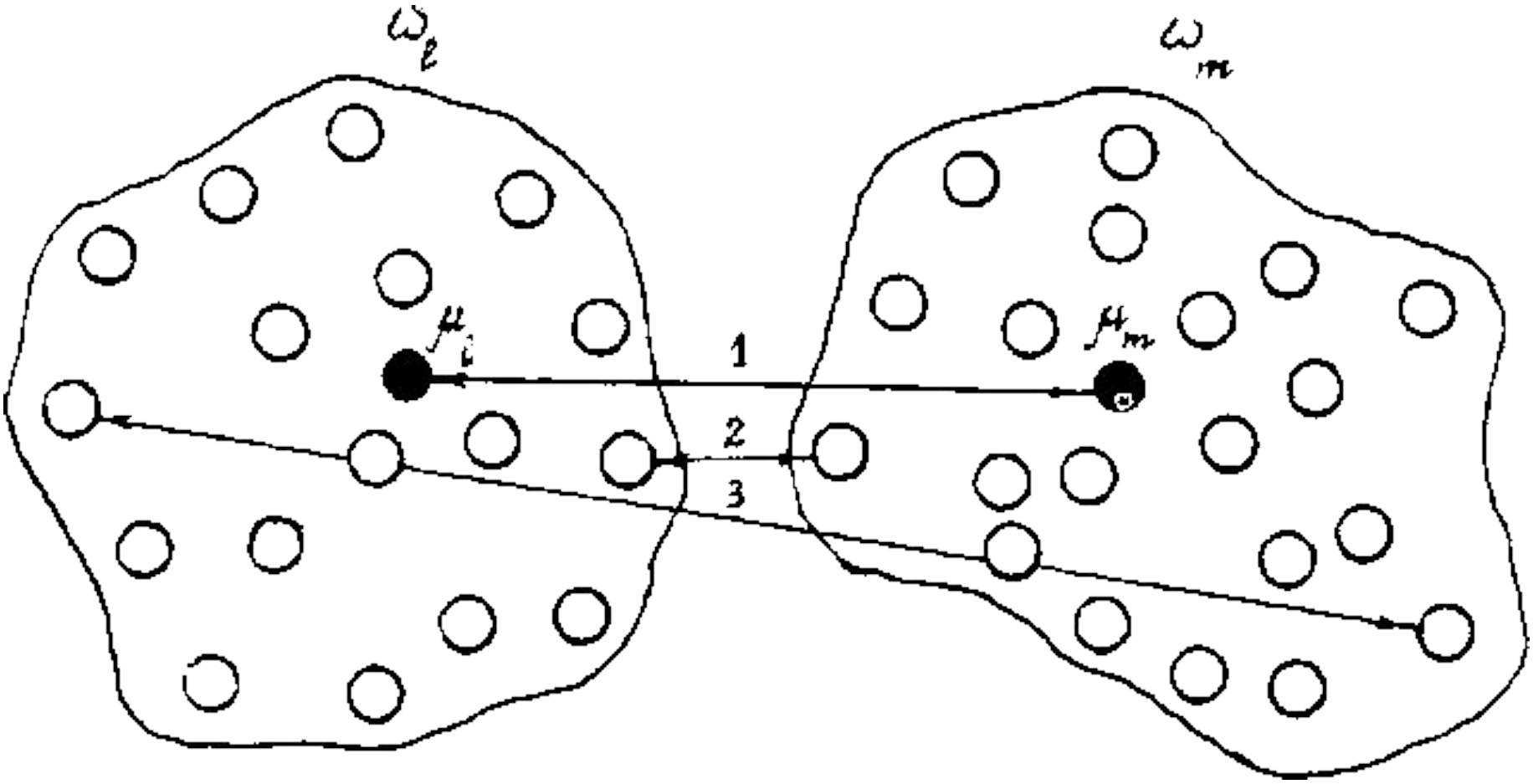

Рис. 1. Различные способы определения расстояния между кластерами w_l и w_m : 1 — по центрам тяжести, 2 — по ближайшим объектам, 3 — по самым далеким объектам

Расстояние ближайшего соседа есть расстояние между ближайшими объектами кластеров:

$$q_{\min}(w_l, w_m) = \min_{x_i \in w_l, x_j \in w_m} d(x_i, x_j) \quad (1)$$

Расстояние дальнего соседа — расстояние между самыми дальними объектами кластеров:

$$q_{\max}(w_l, w_m) = \max_{x_i \in w_l, x_j \in w_m} d(x_i, x_j) \quad (2)$$

Расстояние центров тяжести равно расстоянию между центральными точками кластеров:

$$q(w_l, w_m) = d(\mu_l, \mu_m) \quad (3)$$

Обобщенное (по Колмогорову) расстояние между классами, или обобщенное *K*-расстояние, вычисляется по формуле

$$q_{\tau}^{(K)}(w_l, w_m) = [\frac{1}{N_l N_m} \sum_{x_i \in w_l} \sum_{x_j \in w_m} d^{\tau}(x_i, x_j)]^{\frac{1}{\tau}} \qquad (4)$$

в частности, при $\tau = \infty$ и при $\tau = -\infty$ имеем:

$$q_{\infty}^{(K)}(w_l, w_m) = q_{\max}(w_l, w_m) \qquad (5)$$

$$q_{-\infty}^{(K)}(w_l, w_m) = q_{\min}(w_l, w_m) \qquad (6)$$

Выбор той или иной меры расстояния между кластерами влияет, главным образом, на вид выделяемых алгоритмами кластерного анализа геометрических группировок объектов в пространстве признаков.

Так, алгоритмы, основанные на расстоянии ближайшего соседа, хорошо работают в случае группировок, имеющих сложную, в частности, цепочечную структуру.

Расстояние дальнего соседа применяется, когда искомые группировки образуют в пространстве признаков шаровидные облака.

И промежуточное место занимают алгоритмы, использующие расстояния центров тяжести и средней связи, которые лучше всего работают в случае группировок эллипсоидной формы.

Многообразие алгоритмов кластерного анализа обусловлено также множеством различных критериев, выражающих те или иные аспекты качества автоматического группирования.

Простейший критерий качества непосредственно базируется на величине расстояния между кластерами.

Наиболее часто применяются критерии в виде отношений показателей "населенности" кластеров к расстоянию между ними.

Это, например, может быть отношение суммы межклассовых расстояний к сумме внутриклассовых (между объектами) расстояний или отношение общей дисперсии данных к сумме внутриклассовых дисперсий и дисперсии центров кластеров.