3. Обзор методов кластерного анализа

3.1. Меры сходства

3.2. Агломеративные методы кластерного анализа

3.3. Итеративные методы кластерного анализа

3.4. Критерии качества

3.1. Кластерный анализ заключает в себе богатые возможности для решения многих задач, связанных с классификацией объектов. Он используется в большинстве случаев, когда нет каких-либо априорных гипотез относительно классов общественных явлений.

Сущность кластерного анализа заключается в том, чтобы на основании данных, содержащихся в множестве X, разбить множество объектов N на ряд кластеров (подмножеств) так, чтобы каждый объект N_r принадлежал одному и только одному подмножеству разбиения (кластеру) и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными (несходными)[1].

Пример 1[2]. Некая фирма собирается начать выпуск нового стирального порошка. Разработана анкета, содержащая ряд вопросов, характеризующих отношение респондентов к свойствам продукта. Респонденты должны проранжировать факторы по степени их значимости, начиная с самого важного, – от 1 до 8.

Результаты классификации объектов (респондентов) по переменным (свойствам продукта) представлены в табл. 1.

Таблица 1 – Результаты классификации респондентов по предпочтениям

Свойство продукта	Ранги свойств по сегментам			
	1 (18 %)	2 (7 %)	3 (60 %)	4 (15 %)
Моющая способность	**3**	8	**2**	7
Отдушка	5	5	7	**1**
Цена	8	7	**1**	**2**
Безвредность	**1**	4	8	**3**

[1] Меркушов В.В. Кластерный анализ в исследовании конкурентоспособности регионов / В.В. Меркушов; Федер. агентство по образованию, Самар. гос. экон. акад. – Самара: Изд-во Самар. гос. экон. акад., 2004. – С. 12.

[2] Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: Учеб. пособие / Под ред. И.В. Орловой. – М.: Вузовский учебник, 2009. – С. 91 – 93.

Эффект отбеливания	**2**	6	**3**	6
Подсинивание	4	**3**	6	8
Быстрое растворение	7	**1**	4	5
Отсутствие пыления	6	**2**	5	4

Получилось четыре сегмента, существенно различающиеся между собой по наиболее важным признакам продукта. Эти признаки выделены в таблице. Их можно назвать «сегментообразующими». Легко видеть, что сегмент 3 – самый крупный (60 % от выборки). Это прагматики, для которых важнейшей характеристикой продукта является его цена, а также такие качества, как моющая способность и эффект отбеливания. В следующем по величине сегменте 1, напротив, на первом месте стоит безвредность порошка, цена же занимает последнее место.

Далее может проводиться сегментация по вопросам, касающимся, например, стиля поведения респондентов («покупаю дешевые», «пользуюсь новинками» и т.п.).

Таким образом, результаты кластерного анализа фактически опишут портрет потребителя с рациональной (свойства стирального порошка) и эмоциональной (оценка степени согласия с утверждениями) точек зрения. На основе их можно определить целевую группу качеств, расставить акценты в рекламном сообщении, избавиться от иллюзий относительно исключительности своего товара по какому-либо определенному свойству и т.д.

Большое достоинство кластерного анализа в том, что он позволяет выполнить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ, в отличие от большинства математико-статистических методов, не накладывает никаких ограничений на вид изучаемых объектов и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры рынка, когда показатели весьма разнообразны и затруднительно применение традиционных эконометрических подходов.

Кластерный анализ играет важную роль и для совокупностей временных рядов, характеризующих экономическое развитие. В частности, можно выделить периоды, когда значения соответствующих показателей были достаточно близкими, а также определить группы показателей, динамика которых во времени наиболее схожа.

Необходимость развития и использования методов кластерного анализа продиктована прежде всего тем, что они помогают построить научно обоснованные классификации, выявить внутренние связи между единицами наблюдаемой совокупности. Построение классификаций особенно актуально для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

Методы кластерного анализа могут применяться с целью сжатия информации, в условиях постоянного увеличения и усложнения потоков статистических данных. При этом в задачах социально-экономического прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (с корреляционно-регрессионным, факторным анализом и т.п.).

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения. Так, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникнуть определенные искажения, а также потеряться индивидуальные черты отдельных объектов за счет замены их характеристик обобщенными значениями параметров кластера.

Расстояния между объектами и кластерами

Различия между схемами решения задач классификации во многом определяются тем, что понимают под сходством, однородностью объектов.

Введем вначале такие ключевые для данной главы понятия, как объект и признак.

Под ***объектами*** будем подразумевать конкретные предметы исследования, нуждающиеся в классификации. Такими объектами могут

быть, например, потребители продукции, отличающиеся своими предпочтениями, различные регионы или страны, предприятия, их продукция и т.п.

Признак (синонимы: свойство, переменная, характеристика) представляет собой конкретное свойство объекта.

Различные свойства могут выражаться как *числовыми*, так и нечисловыми значениями.

Обычной формой представления исходных данных в задачах кластерного анализа служит прямоугольная таблица «объект – признак»

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{pmatrix},$$

каждая строка которой представляет результат измерений m рассматриваемых признаков на одном из n обследованных объектов (таблица 2).

Каждая единица статистической совокупности в кластерном анализе рассматривается как точка в заданном признаковом многомерном пространстве. Значение каждого из признаков у данной единицы служит ее координатой в этом пространстве.

Пример 2.

Таблица 2. – Частные показатели эффективности добычи полезных ископаемых субъектов Сибирского федерального округа в 2011 г.

N п/п	**Показатели**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**
	Регион	**2832,6**	**0,650**	**1,15**	**0,21**	**47,3**
1	Республика Алтай	558,9	0,595	3,49	0,11	25
2	Республика Бурятия	966,7	0,759	0,81	0,46	50,5
3	Республика Тыва	557,1	0,609	1,76	0,76	1,3
4	Республика Хакасия	1769,5	0,675	2,80	0,08	51,8
5	Алтайский край	1152,6	0,695	1,14	0,27	37,4
6	Забайкальский край	1257,7	0,692	2,78	0,07	9,5
7	Красноярский край	6766,3	0,849	0,97	0,31	99
8	Иркутская область	3438,9	0,604	1,11	0,26	118,1
9	Кемеровская область	1938,8	0,519	1,59	0,15	32,3
10	Новосибирская область	2205,3	0,778	0,81	0,17	32,8

11	Омская область	6281,6	0,406	0,76	0,08	-27,7
12	Томская область	8722,0	0,783	0,60	0,26	17,2

Показатели эффективности:

x_1 – производительность труда – валовой региональный продукт в расчете на одного занятого;

x_2 – продуктивность – валовая добавленная стоимость на рубль выпущенной продукции;

x_3 – фондоотдача – объем реализованной продукции на рубль основных фондов;

x_4 – инвестиционная отдача – объем инвестиций в основной капитал на один рубль выпущенной продукции;

x_5 – рентабельность (убыточность) проданных товаров, продукции (работ, услуг) организаций – отношение прибыли (убытка) от продаж к затратам на производство проданных товаров, продукции, работ, услуг (рассчитывается Росстатом).

Основой кластерного анализа является определение меры однородности (близости) объектов. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние.

Расстоянием между i-м и j-м объектами в пространстве признаков называется такая величина d_{ij}, которая удовлетворяет следующим аксиомам:

1. $d_{ij} \geq 0$ (неотрицательность);
2. $d_{ij} = d_{ji}$ (симметрия);
3. $d_{ij} + d_{jq} \geq d_{iq}$ (неравенство треугольника, здесь q – номер объекта);
4. если $d_{ij} \neq 0$, то $i \neq j$ (различимость нетождественных объектов);
5. если $d_{ij} = 0$, то $i = j$ (неразличимость тождественных объектов)[3].

В кластерном анализе используются различные меры расстояния между объектами:

[3] Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: Учеб. пособие / Под ред. И.В. Орловой. – М.: Вузовский учебник, 2009. – С. 96.

евклидово расстояние: $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$, одно из наиболее известных расстояний, которое доступно для восприятия и понимания в случае количественных признаков (как и квадрат расстояния).

квадрат евклидова расстояния: $d_{ij} = (\sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2})^2$,

взвешенное евклидово расстояние: $d_{ij} = \sqrt{\sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2}$, весовые коэффициенты которого придают отдельным слагаемым в сумме большую значимость.

расстояние Минковского $d_{ij} = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p \right)^{1/p}$

манхеттенское расстояние (Хеммингово расстояние, «расстояние городских кварталов», citi-block): $d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$, частный случай расстояния Минковского, широко используется для дихотомических качественных признаков, относящихся к номинальной шкале.

расстояние Махалонобиса: $d_{ij} = (X_i - X_j) S_*^{-1} (X_i - X_j)$,

где d_{ij} – расстояние между объектами i и j; x_{ik}, x_{jk} – значения k-ой переменной для i-го и j-го объекта, соответственно; X_i, X_j – векторы значений переменных у i-го и j-го объектов; S_* – общая ковариационная матрица, w_k – вес, приписываемый k-ой переменной; p – показатель степени, определяемый исследователем[4].

Оценка сходства между объектами сильно зависит от абсолютного значения признака и от степени его вариации в совокупности. Чтобы устранить подобное влияние на процедуру классификации, можно значения

[4] Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: Учеб. пособие / Под ред. И.В. Орловой. – М.: Вузовский учебник, 2009. – С. 96–97.; Многомерный статистический анализ в экономике: Учеб. пособие для вузов / Под ред. проф. В.Н. Томашевича. – М.: ЮНИТИ-ДАНА, 1999. – С. 498–471.

исходных переменных нормировать. Можно выделить ряд способов стандартизации (нормирования) показателей:

1) $Y_{rj} = \frac{X_{rj} - \overline{X_j}}{S_{X_j}}$, где X_{rj} –значение j-го показателя для r-го объекта, $\overline{X_j}$ – среднее, S_{X_j} – дисперсия[5];

При таком подходе также возникают проблемы, поскольку значения Y_{rj} могут быть как положительными, так и отрицательными.

2) $Y_{rj} = \frac{Z_{rj} - \min Z_{rj}}{\max Z_{rj} - \min Z_{rj}}$, где $Z_{rj} = X_{rj} - X_{kj}^0$ (для показателей-стимулянтов); $Z_{rj} = X_{kj}^0 - X_{rj}$ (для показателей-дестимулянтов), где X_{kj}^0 – значение j-го показателя для объекта k, принятого за образец ($k \neq 1$)[6]:

3) $Y_{rj} = \frac{X_{rj}}{X_{max}}$. Метод «Паттерн» позволяет получить оценки по частным показателям при помощи соотнесения фактических значений с наилучшими[7].

Пример 3.

Таблица 3. – Стандартизированные частные показатели эффективности добычи полезных ископаемых субъектов Сибирского федерального округа в 2011 г.

N п/п	**Показатели**	Z_1	Z_2	Z_3	Z_4	Z_5
	Регион					
1	Республика Алтай	0,064	0,701	1,000	0,150	0,212
2	Республика Бурятия	0,111	0,893	0,231	0,606	0,428
3	Республика Тыва	0,064	0,717	0,504	1,000	0,011
4	Республика Хакасия	0,203	0,794	0,802	0,103	0,439
5	Алтайский край	0,132	0,818	0,327	0,362	0,317
6	Забайкальский край	0,144	0,815	0,795	0,092	0,080
7	Красноярский край	0,776	1,000	0,276	0,408	0,838
8	Иркутская область	0,394	0,711	0,318	0,349	1,000
9	Кемеровская область	0,222	0,611	0,455	0,193	0,273
10	Новосибирская область	0,253	0,916	0,233	0,227	0,278

[5] Евченко А.В. Исследование и регулирование регионального развития с использованием комплексных социально-экономических индикаторов: Монография / Курск. гос. техн. ун-т. Курск, 2004. – С. 93.

[6] Евченко А.В. Исследование и регулирование регионального развития с использованием комплексных социально-экономических индикаторов: Монография / Курск. гос. техн. ун-т. Курск, 2004. – С. 92–93.

[7] Белякова Е.В. Методы оценки регионального развития в условиях глобализации: монограф. / Е.В. Белякова, Н.В. Веретенов; Сиб. гос. аэрокосмич. ун-т. – Красноярск, 2006. – С. 47.

11	Омская область	0,720	0,478	0,219	0,100	-0,235
12	Томская область	1,000	0,922	0,173	0,343	0,146

Вопрос о придании переменным соответствующих весов должен решаться после проведения исследователем тщательного анализа изучаемой совокупности и социально-экономической сущности классифицируемых переменных.

Если алгоритм кластеризации основан на измерении сходства между переменными, то в качестве мер сходства могут быть использованы также линейные коэффициенты корреляции, коэффициенты ранговой корреляции и т.д.

3.2. Большинство методов иерархической кластеризации являются агломеративными (от латинского agglomero – присоединяю, накапливаю) – процесс начинается с создания элементарных кластеров, каждый из которых состоит ровно из одного исходного наблюдения (одной точки), а на каждом последующем шаге происходит объединение двух наиболее близких кластеров в один.

Отдельные методы кластерного анализа различаются тем, что в них по-разному выбирается способ определения близости между кластерами (и между объектами), используются различные алгоритмы вычислений и меры расстояний между кластерами (объектами). Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.

Графическое изображение процесса объединения кластеров может быть получено с помощью дендрограммы – дерева объединения кластеров[8]. На дендрограмме указываются номера объединяемых объектов и расстояние (или иная мера сходства), при котором произошло объединение.

Наиболее часто используемыми агломеративными (иерархическими) методами являются:

•метод «ближнего соседа» (метод одиночной связи – «single linkage»);

Пример 4.

[8] Дубров А.М., Мхиторян В.С., Трошин Л.И. Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2000. – С. 200.

Таблица 4 – Матрица Евклидовых расстояний между объектами

Регион	РА	РБ	РТ	РХ	АК	ЗК	КК	ИО	КО	НО	ОО	ТО
РА	,000	,941	1,005	,348	,726	,287	1,257	1,111	,579	,825	1,131	1,285
РБ	,941	,000	,661	,773	,295	,843	,814	,717	,580	,432	1,096	,971
РТ	1,005	,661	,000	1,050	,739	,964	1,294	1,243	,871	,904	1,197	1,216
РХ	,348	,773	1,050	,000	,560	,365	,949	,808	,436	,618	1,064	1,091
АК	,726	,295	,739	,560	,000	,590	,851	,739	,313	,230	,904	,904
ЗК	,287	,843	,964	,365	,590	,000	1,174	1,101	,459	,629	,934	1,095
КК	1,257	,814	1,294	,949	,851	1,174	,000	,511	,925	,793	1,209	,742
ИО	1,111	,717	1,243	,808	,739	1,101	,511	,000	,782	,778	1,295	1,078
КО	,579	,580	,871	,436	,313	,459	,925	,782	,000	,380	,749	,906
НО	,825	,432	,904	,618	,230	,629	,793	,778	,380	,000	,814	,770
ОО	1,131	1,096	1,197	1,064	,904	,934	1,209	1,295	,749	,814	,000	,680
ТО	1,285	,971	1,216	1,091	,904	1,095	,742	1,078	,906	,770	,680	,000

В первый кластер будут включены Алтайский край и Новосибирская область, так как расстояние между ними минимальное (d = 0,230). Поскольку есть регионы со схожими мерами близости, параллельно идет образование следующего кластера (назовем его вторым), объединяющего Республику Алтай и Забайкальский край (d = 0,287). На следующем шаге к первому кластеру будет подключена Республика Бурятия, так как расстояние она к выделенным регионам ближе всех (d = 0,295). Затем к ним добавляется Кемеровская область (d = 0,313). Ко второму кластеру добавляется Республика Хакасия (d = 0,348). И так до конца. Графически это будет выглядеть следующим образом.

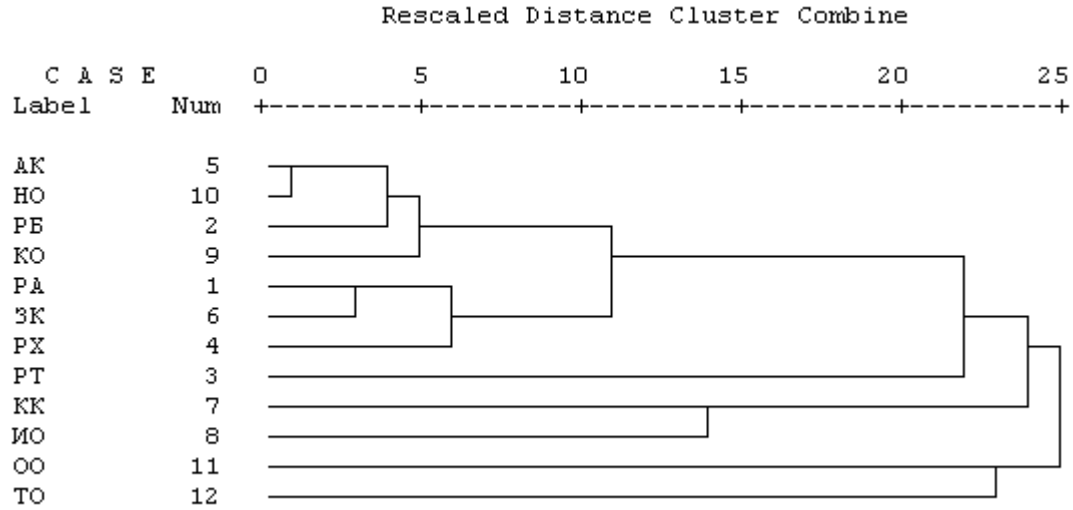
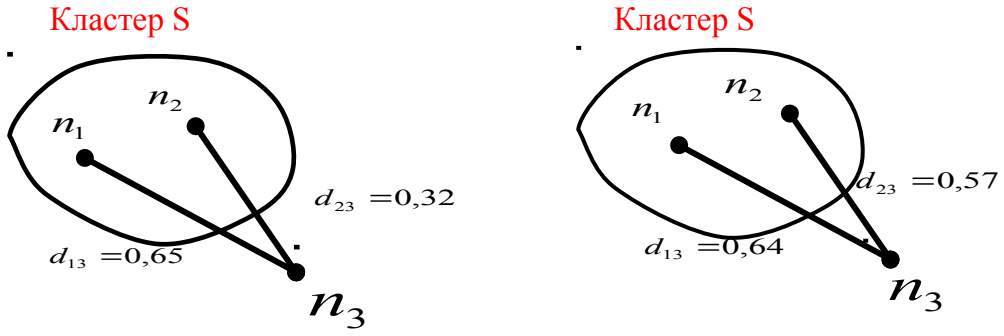


Рисунок 1 – Дендрограмма кластеризации регионов СФО, построенная методом одиночной связи (пример 3)

•метод «дальнего соседа» (метод полной связи – «complete linkage»);



а) если задано предельное расстояние 0,3, то третий объект не будет включен в кластер S

б) если задано предельное расстояние 0,7, то третий объект будет включен в кластер S

Рисунок 2 – Определение состава кластера при различных уровнях сходства наблюдаемых объектов

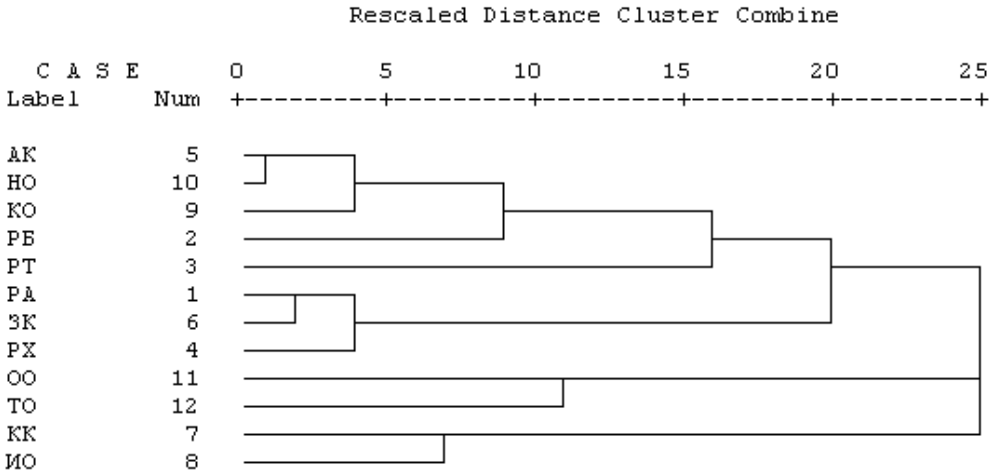


Рисунок 3 – Дендрограмма кластеризации регионов СФО, построенная методом полной связи (пример 3)

•метод «средней связи» («pair-group average»);

Программный пакет SPSS предлагает 4 метода объединения, основанных на средних значениях[9]:

1) связь между группами (between-groups linkage): расстояние между кластерами равно среднему значению расстояния между всеми возможными парами наблюдений, причем одно наблюдение берется из одного кластера, другое – из другого. Информация, необходимая для расчета дистанции, находится на основании всех теоретически возможных пар наблюдений. По этой причине в программе он устанавливается по умолчанию.

[9] Брююль А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Брююль, П. Цёфель. – СПб.: ООО «ДиаСофртЮП», 2005. – С. 402. (608 с)

2) связь внутри группы (within-groups linkage). Расстояние между двумя кластерами рассчитывается на основании всех возможных пар наблюдений, принадлежащих обоим кластерам, причем учитываются также и пары наблюдений, образующиеся внутри кластера.

3) центроидная кластеризация (centroid clustering): в обоих кластерах рассчитываются средние значения переменных, относящихся к ним наблюдений. Затем расстояние между двумя кластерами рассчитывается как расстояние между усредненными наблюдениями. Центроид нового кластера получается как взвешенное среднее центроидов обоих исходных кластеров, причем количество наблюдений исходных кластеров образуют весовой коэффициент;

4) медианная кластеризация (median clustering). При определении среднего центроида кластеров, оба кластера берутся с одинаковыми весами.

Геометрический пример с двумя кластерами представлен ниже.

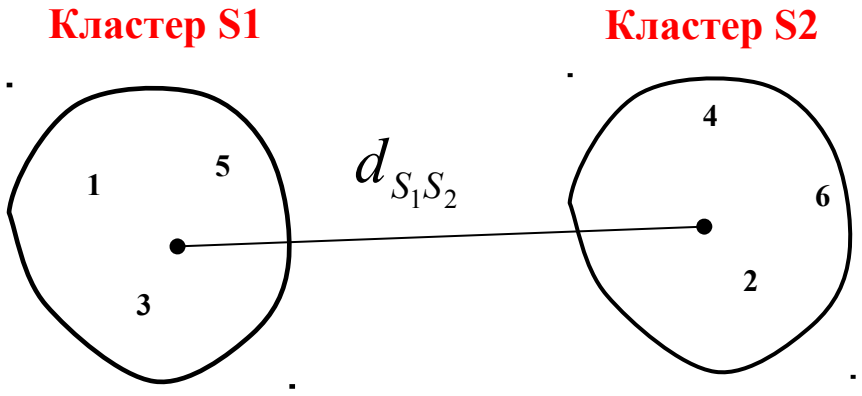


Рисунок 4 – Объединение двух кластеров по методу средней связи

Каждый кластер содержит по три объекта. Чтобы решить вопрос об объединении этих кластеров, определяют их центры тяжести и расстояние между ними. Если расстояние между ними меньше заданного уровня, кластеры объединятся в один.

•метод Уорда («Ward's method»).

Данный метод предполагает, что первоначально каждый кластер состоит из одного объекта. Сначала объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений V_k:

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left(x_{ij} - \bar{x}_{jk}\right)^2,$$

где k – номер кластера; i – номер объекта; j – номер признака; p – количество признаков, характеризующих каждый объект; n_k – количество объектов в k-ом кластере[10].

Пример 5. Алгоритм работы методы Уорда[11]. Пусть имеются пять объектов, каждый из которых характеризуется двумя признаками.

Таблица 5 – Матрица исходных данных

Номер объекта	X_1	X_2
1	21	0,3
2	18	0,8
3	15	0,6
4	13	0,7
8	25	0,9

Рассчитаем матрицу расстояний с использованием евклидовой метрики:

$$\Delta = \begin{pmatrix} 0 & 3{,}0141 & 6{,}0075 & 8{,}0099 & 4{,}0447 \\ & 0 & 3{,}0066 & 5{,}0010 & 7{,}0007 \\ & & 0 & 2{,}0025 & 10{,}0045 \\ & & & 0 & 12{,}0017 \\ & & & & 0 \end{pmatrix}.$$

Самые близкие объекты n_3 и n_4 объединяются в один кластер. Для нового кластера определяем сумму квадратов отклонений V_k.в данном случае присвоим этому кластеру номер S_3, т.е. k=3, и рассчитаем величину V_k:

$$V_k = \sum_{i=1}^{2} \sum_{j=1}^{2} \left(x_{ij} - \bar{x}_{j3}\right)^2,$$

где $\bar{x}_{j3}$ – среднее значение j-го признака в кластере S_3 ($\bar{x}_{13} = 14$, $\bar{x}_{23} = 0{,}65$).

$$V_3 = \left[(15 - 14)^2 + (0{,}6 - 0{,}65)^2\right] + \left[(13 - 14)^2 + (0{,}7 - 0{,}65)^2\right] = 2{,}005.$$

Теперь нужно решить вопрос о том, какой новый объект может быть на следующем шаге присоединен к третьему кластеру или какие кластеры можно объединить. Если объединить первый и второй объекты, тогда V = 4,625; для первого и пятого V = 8,0081, для второго и пятого V = 12,25.

Теперь попробуем присоединять к кластеру S_3 поочередно каждый из оставшихся объектов. Если к S_3 добавить n_1, то V = 34,75; для S_3 и n_2 V =

[10] Киреев А.В. Теоретические аспекты цифровой обработки многомерных данных и сигналов: монография. – Пенза: Приволжский Дом знаний, 2012. – С. 47.

[11] Многомерный статистический анализ в экономике: Учеб. пособие для вузов / Под ред. проф. В.Н. Томашевича. – М.: ЮНИТИ-ДАНА, 1999. – С. 477–479.

12,69; для S_3 и n_5 V = 82,71. очевидно, что наименьшее отклонение дает объединение объектов n_1 и n_2 в отдельный кластер (S_1).

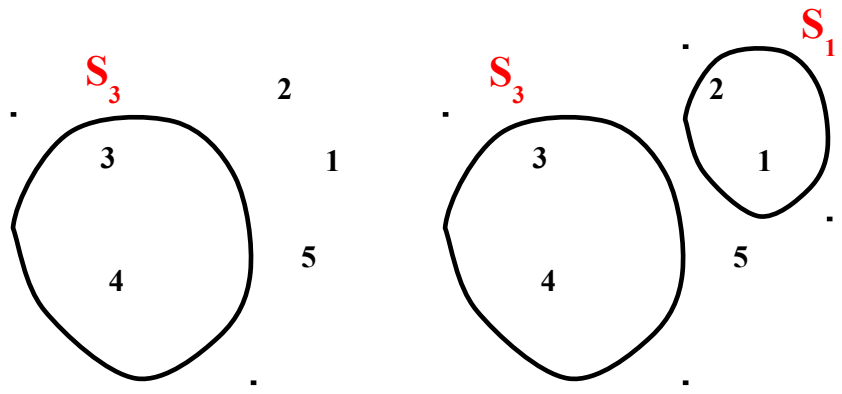


Рисунок 5– Кластеризация объектов методом Уорда

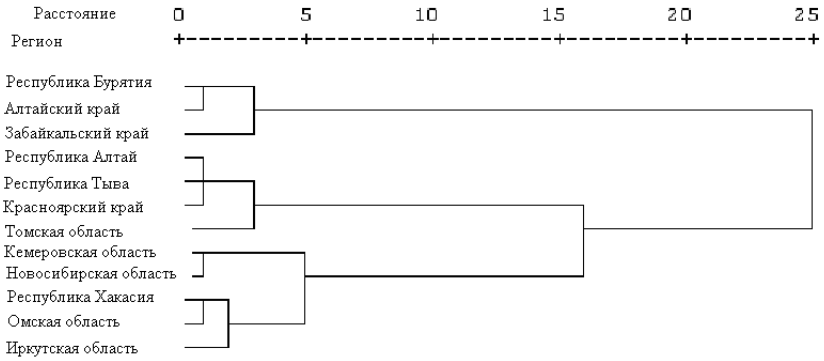


Рисунок 6 – Дендрограмма кластеризации регионов СФО, построенная методом Уорда (пример).

Алгоритм иерархического кластерного анализа можно представить в виде последовательности процедур:

Шаг 1. Нормирование значений исходных переменных.

Шаг 2. Расчет матрицы расстояний или матрицы мер сходства.

Шаг 3. Поиск пары самых близких кластеров. По выбранному алгоритму объединяются эти два кластера. Новому кластеру присваивается меньший из номеров объединяемых кластеров.

Шаг 4. Повторение шага 2 и 3 до объединения всех объектов в один кластер или до достижения заданного «порога» сходства.

Кроме рассмотренных агломеративных методов иерархического кластерного анализа, существуют методы, противоположные им по логическому построению процедур классификации. Они называются иерархические дивизимные методы. Основной исходной посылкой дивизимных методов является то, что первоначально все объекты принадлежат одному кластеру. В процессе классификации по опредленным правилам постепенно от этого кластера отделяются группы схожих между собой объектов. Таким образом, на каждом шаге количество кластеров возрастает, а мера рассеяния между кластерами уменьшается. Дендрограмма для дивизимных методов представлена ниже.

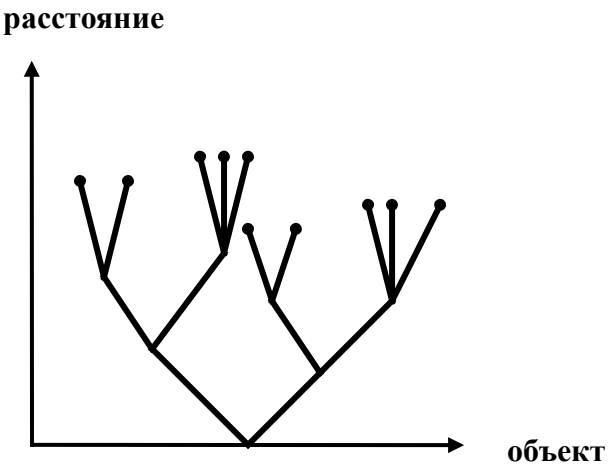


Рисунок 7 – Дендрограмма кластеризации, построенная для дивизимных методов (пример)

Пример 6. Пусть дана матрица расстояний между объектами

$$\Delta = \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 & 3{,}24 \\ & 0 & 3{,}26 & 1{,}92 & 1{,}93 \\ & & 0 & 2{,}68 & 2{,}74 \\ & & & 0 & 0{,}71 \\ & & & & 0 \end{pmatrix}$$

Требуется провести классификацию по дивизимному алгоритму.

Решение. Наиболее удаленными являются объекты n_1 и n_2 (d_{12} = 4,49); оценим расстояния оставшихся объектов до первого и второго:

$d_{31} < d_{32}$ объект n_3 ближе к n_1;

$d_{41} > d_{42}$ объект n_4 ближе к n_2;

$d_{51} > d_{52}$ объект n_5 ближе к n_2;

Таким образом, получаем два кластера: $S_1\{1, 3\}$ и $S_2\{2, 4, 5\}$. В каждом из них анализируем расстояния между объектами, и на очередном шаге происходит разделение того кластера, где достигается максимум расстояния между объектами:

$d_{13} = 2,16; d_{25} = 1,93; d_{14} = 1,92; d_{45} = 0,71.$

Наибольшее расстояние $d_{13} = 2,16$, следовательно, объекты n_1 и n_3 выделяем в отдельные кластеры. В кластере $S_2\{2, 4, 5\}$ ищем максимальное расстояние $\max\{d_{24}, d_{25}, d_{45}\} = 1,93$. На следующем шаге из этого кластера выделяем объект n_2 и, наконец, на последнем шаге разделяем кластер $S_4\{4, 5\}$ на два кластера на расстоянии 0,71.

Дендрограмма последовательности разбиений представлена на рис. 8.

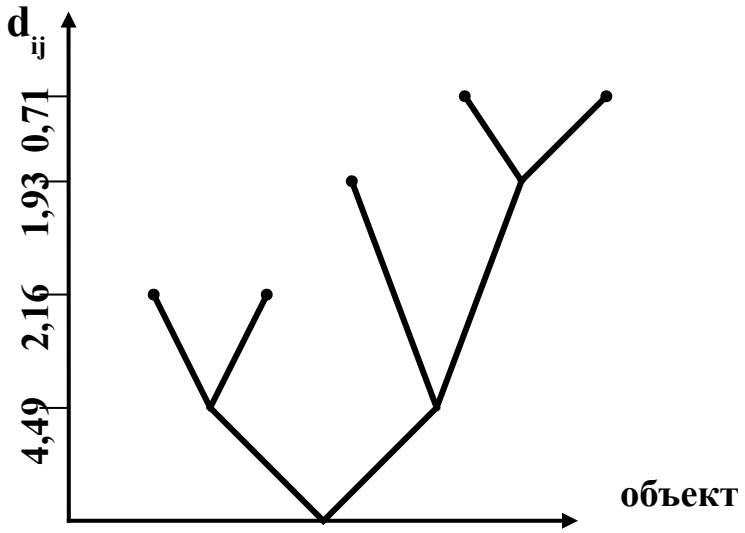


Рисунок 8 – Дендрограмма кластеризации

Из этого примера видно, что дивизимный алгоритм не требует пересчета матрицы расстояний на каждом шаге классификации, в отличие от агломеративных методов.

Иерархические методы используются обычно в таких задачах классификации небольшого числа объектов (порядка нескольких десятков), где больший интерес представляет не число кластеров, а анализ структуры множества этих объектов и наглядная интерпретация проведенного анализа в виде дендрограммы.

3.3. Наряду с иерархическими методами классификации группа итеративных методов кластерного анализа. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.). Итеративные методы в большей степени, чем иерархические, требуют от пользователя интуиции при выборе типа классификационных процедур и задания начальных условий разбиения, так как большинство этих методов очень чувствительны к изменению задаваемых параметров. Например, выбранное случайным образом число кластеров может не только сильно увеличить трудоемкость процесса классификации, но и привести к образованию «размытых» или мало наполняемых кластеров. Поэтому целесообразно сначала провести классификацию по одному из иерархических методов или на основании экспертных оценок, а затем уже подбирать начальное разбиение и статистический критерий для работы итерационного алгоритма. Как и в иерархическом кластерном анализе, в итерационных методах существует проблема определения числа кластеров. В общем случае их число может быть неизвестно. Не все итеративные методы требуют первоначального задания числа кластеров. Но для окончательного решения вопроса о структуре изучаемой совокупности можно испробовать несколько алгоритмов, меняя либо число образуемых кластеров, либо установленный порог близости для объединения объектов в кластеры. Тогда появляется возможность выбрать наилучшее разбиение по задаваемому критерию качества.

Метод k-средних принадлежит к группе итеративных методов эталонного типа. Само название метода было предложено Дж. Мак-Куином в 1967 г.[12]

В отличие от иерархических процедур метод k-средних не требует вычисления и хранения матрицы расстояний или сходств между объектами.

[12] Многомерный статистический анализ в экономике: Учеб. пособие для вузов / Под ред. проф. В.Н. Томашевича. – М.: ЮНИТИ-ДАНА, 1999. – С. 486–493.

Алгоритм этого метода предполагает использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы k случайно выбранных объектов, которые будут служить эталонами, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие. В этом случае важную роль играет выбор начальных условий, которые влияют на длительность процесса классификации и на его результаты.

Метод k-средних удобен для обработки больших статистических совокупностей.

Рассмотрим математическое описание алгоритма метода.

Пусть имеется n наблюдений, каждое из которых характеризуется p признаками $X_1, X_2, \ldots, X_p$. Эти наблюдения необходимо разбить на k кластеров. Для начала из n точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо априорных соображений k точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера. На первом шаге из оставшихся $(n - k)$ объектов извлекается точка X_i с координатами $(x_{i1}, x_{i2}, \ldots, x_{ip})$ и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например, евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{lj})^2}$$

Проверяемый объект присоединяется к тому центру (эталону), которому соответствует $min\ d_{il}\ (l = 1, \ldots, k)$. Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более минимальных расстояния, то i-й объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбираем точку X_{i+1} и для нее повторяются все процедуры. Таким образом, через (n – k) шагов все точки (объекты) совокупности окажутся

отнесенными к одному из k кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки X_1, X_2,…, X_n опять подсоединяются к полученным кластерам, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить C_1, C_2, …, C_k. При этом каждая точка X_i (i =1, 2, …, n) будет относиться к такому кластеру (классу) l, для которого:

$$d(x_j, c_l) = \min_{1 \le j \le R} d(x_j, c_j)$$

Возможны две модификации метода k-средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется. Рассмотрим работу итеративного алгоритма метода k-средних на примере.

Пример 7. Пусть имеются шесть объектов, которые необходимо разбить на три класса (кластера) при помощи метода k-средних. Каждый из объектов описывается тремя переменными X_1, X_2, X_3. Исходные значения этих переменных представлены в табл. 4.

Таблица 4 – Исходные данные

Номер объекта	Y_1	Y_2	Y_3
1	0,10	10	5,0
2	0,80	14	2,0
3	0,40	12	3,0
4	0,18	11	4,0
5	0,25	13	3,2
6	0,67	15	2,4

В качестве эталонов возьмем первые три объекта (k = 3). Согласно выбранному правилу классификации запишем исходные значения эталонов и весов:

$$E_1^0 = X_1 = (0{,}10;10;5{,}0),\ w_1^0 = 1$$
$$E_2^0 = X_2 = (0{,}80;14;2{,}0),\ w_2^0 = 1 \quad \text{- нулевая итерация}$$
$$E_3^0 = X_3 = (0{,}40;12;3{,}0),\ w_3^0 = 1$$

На *первом шаге* берем четвертый объект и определяем его расстояние до каждого из эталонов по евклидовой метрике:

$$d_{41} = \sqrt{(0{,}18 - 0{,}10)^2 + (11 - 10)^2 + (4{,}0 - 5{,}0)^2} = 1{,}416,$$
$$d_{42} = \sqrt{(0{,}18 - 0{,}80)^2 + (11 - 14)^2 + (4{,}0 - 2{,}0)^2} = 3{,}222,$$
$$d_{43} = \sqrt{(0{,}18 - 0{,}40)^2 + (11 - 12)^2 + (4{,}0 - 3{,}0)^2} = 1{,}432$$

Следовательно, рассматриваемый объект должен быть присоединен к первому эталону и первый эталон будет пересчитан, а второй и третий не меняются:

$$E_1^1 = \frac{w_1^0 \cdot E_1^0 + X_4}{w_1^0 + 1},\ w_1^1 = w_1^0 + 1 = 2,\ E_2^1 = E_2^0,\ E_3^1 = E_3^0,\ w_2^1 = w_2^0,\ w_3^1 = w_3^0,$$

где X_4 – вектор значений переменных для четвертого объекта.

$$E_1^1 = \left(\frac{0{,}10 + 0{,}18}{2}; \frac{10 + 11}{2}; \frac{5 + 4}{2}\right) = (0{,}14;\ 10{,}5;\ 4{,}5)$$

На *втором шаге* проверяем, к какому эталону ближе всего находится пятый объект:

$$d_{51} = \sqrt{(0{,}25 - 0{,}14)^2 + (13 - 10{,}5)^2 + (3{,}2 - 4{,}5)^2} = 2{,}820,$$
$$d_{52} = \sqrt{(0{,}25 - 0{,}80)^2 + (13 - 14)^2 + (3{,}2 - 2{,}0)^2} = 1{,}656,$$
$$d_{53} = \sqrt{(0{,}25 - 0{,}40)^2 + (13 - 12)^2 + (3{,}2 - 3{,}0)^2} = 1{,}031$$

Так как $d_{53} = \min(d_{51}, d_{52}, d_{53})$, следовательно, пятый объект присоединяется к третьему эталону, этот эталон пересчитывается и вес его увеличивается:

$$E_3^2 = \left(\frac{0{,}40 + 0{,}25}{2}; \frac{12 + 12}{2}; \frac{3 + 3{,}2}{2}\right) = (0{,}325;\ 12{,}5;\ 3{,}1)$$

$$w_3^2 = w_3^1 + 1 = 2,\ E_2^2 = E_2^1,\ E_1^2 = E_1^1,\ w_1^2 = w_1^1,\ w_2^2 = w_2^1$$

На *третьем шаге* все рассуждения повторяем для шестого объекта:

$$d_{61} = \sqrt{(0{,}67 - 0{,}14)^2 + (15 - 10{,}5)^2 + (2{,}4 - 4{,}5)^2} = 4{,}994,$$
$$d_{62} = \sqrt{(0{,}67 - 0{,}80)^2 + (15 - 14)^2 + (2{,}4 - 2{,}0)^2} = 1{,}085,$$
$$d_{63} = \sqrt{(0{,}67 - 0{,}325)^2 + (15 - 12{,}5)^2 + (2{,}4 - 3{,}1)^2} = 2{,}619$$

Пересчитываем второй эталон и его вес:

$$E_2^3 = \left(\frac{0{,}80 + 0{,}67}{2}; \frac{14 + 15}{2}; \frac{2{,}4 + 2{,}0}{2} \right) = (0{,}735;\ 14{,}5;\ 2{,}2)$$

$$w_2^3 = w_2^2 + 1 = 2,\ E_1^3 = E_1^2,\ E_3^3 = E_3^2,\ w_1^3 = w_1^2,\ w_3^3 = w_3^2$$

После того как просмотрены все объекты, кроме первых трех, процесс «зацикливается», т.е. по тому же правилу осуществляются просмотр и присоединение к соответствующему эталону каждого из шести объектов. При этом происходит пересчет эталонов и продолжается наращивание их весов. Процесс завершается тогда, когда последующее разбиение (итерации 16–21) дали такой же результат, как и предыдущее разбиение (итерации 10–15).

Образованы три кластера: S_1 {1}, S_2 {2, 6}, S_3 {3, 4, 5}. Вычисляем центры тяжести полученных кластеров, причем в общем случае эти центры не совпадают с эталонами:

C_1 = (0,10; 10; 5,0) – центр 1 кластера,

C2 = (0,735; 14,5; 2,2) – центр 2 кластера,

C_3 = (0,277; 12,00; 3,4) – центр 3 кластера.

После этого строится окончательное разбиение: каждая многомерная точка относится к тому кластеру, центр которого ближе всех к этой точке.

Для нашего примера определяем поочередно расстояния всех точек (X_1, X_2, X_3, X_4, X_5, X_6) до центров трех кластеров (табл. 5).

Таблица 5 – Расстояния до центров классов

Центры кластеров	Объекты					
	1	2	3	4	5	6
C_1	0	5,049	2,844	1,416	3,502	5,664
C_2	5,338	0,542	2,646	3,939	1,867	0,542
C_3	5,920	2,497	0,418	1,169	1,020	3,187

Как видно из табл. 5, подтверждается полученное разбиение на три кластера: S_1 {1}, S_2 {2, 6}, S_3 {3, 4, 5}. На этом алгоритм завершается.

Если было проведено несколько разбиений заданной совокупности, то необходимо оценить качество каждого разбиения на основании одного из выбранных критериев, чтобы прийти к окончательному решению.

Рассмотренный метод k-средних допускает в качестве исходного разбиения использовать группировку, полученную одним из методов иерархического кластерного анализа. Такой подход можно рекомендовать для сокращения времени обработки в том случае, когда совокупность объектов достаточно велика и пользователь затрудняется указать количество образуемых кластеров.

Вычислительные процедуры большинства итеративных методов классификации сводятся к выполнению следующих шагов:

Шаг 1. Выбор числа кластеров, на которые должна быть разбита совокупность, задание первоначального разбиения объектов и определение центров тяжести кластеров.

Шаг 2. В соответствии с выбранными мерами сходства определение нового состава каждого кластера.

Шаг 3. После полного просмотра всех объектов и распределения их по кластерам осуществляется пересчет центров тяжести кластеров.

Шаг 4. Процедуры 2 и 3 повторяются до тех пор, пока следующая итерация не даст такой же состав кластеров, что и предыдущая.

Аналогично работает *параллельный пороговый метод (parallel threshold method)*, за исключением того, что одновременно выбирают несколько кластерных центров и объекты в пределах порогового уровня группируют с ближайшим центром.

Метод оптимизирующего распределения (optimizing partitioning method) отличается от двух изложенных выше пороговых методов тем, что объекты можно впоследствии поставить в соответствие другим кластерам (перераспределить), чтобы оптимизировать суммарный критерий, такой как среднее внутри кластерное расстояние для данного числа кластеров.

Одним из итеративных методов классификации, не требующих задания числа кластеров, является *метод поиска сгущений*. В теории и на практике существует несколько различных модификаций этого метода. Каждая модификация отличается задаваемым начальным состоянием и критериями завершения классификации. Остановимся подробно на одном из алгоритмов поиска сгущений, который получил название «форель». Суть итеративного алгоритма типа «форель» заключается в применении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков с целью поиска локальных сгущений точек. Рассмотрим схему данного алгоритма в общем виде и на конкретном примере.

Существует очень схожий с ним алгоритм взаимного поглощения, который также является итерационным и использует идею гиперсферы. Суть его заключается в том, что для каждой многомерной величины X_i определяется свой радиус R_i, например следующим образом:

$$R_i = \max_l d(X_i, X_l) - \delta$$

где $d(X_i, X_l)$ – расстояние от i-й до l-й точки; δ – некоторая выбираемая величина, постоянная для всех точек.

Сферы с радиусами R, строятся с центрами в точках X_i (i = 1, …, n). Области пересечения, содержащие центры этих сфер, называются областью взаимного поглощения. А совокупность центров сфер, попавших в эту область, называется кластером.

Графически для двумерных величин это можно представить так, как показано на рис 5.

Области взаимного поглощения на рисунке заштрихованы и ясно видны группы точек, попавших в два кластера S_1 и S_2. Изменяя величину δ, можно повторить разбиение несколько раз. В качестве окончательного решения задачи следует выбирать вариант разбиения, сохраняющийся при нескольких значениях δ, как наиболее устойчивый.

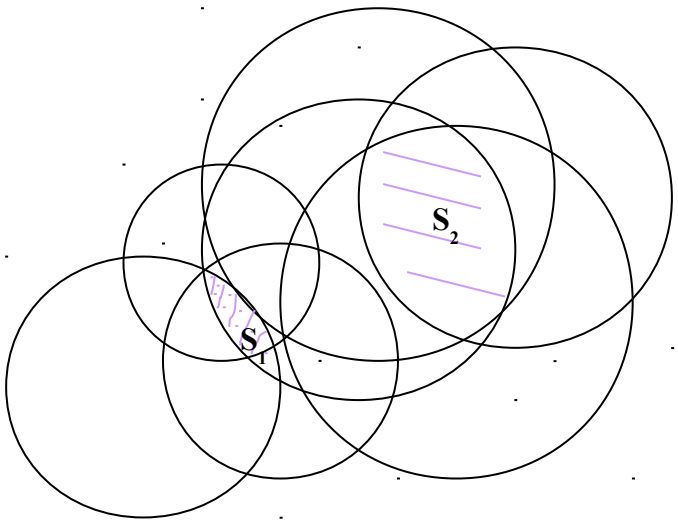


Рисунок 9 – Области взаимного поглощения и два образованных ими кластера

Итак, еще раз подчеркнем некоторые особенности методов:

1. В качестве метрики используется Евклидово расстояние
2. Число кластеров заранее не известно и выбирается исследователем заранее
3. Качество кластеризации зависит от первоначального разбиения

3.4. Критерии качества классификации

При использовании различных методов кластерного анализа для одной и той же совокупности могут быть получены различные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают: во-первых, набор признаков, по которым осуществляется классификация, во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу, и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Если используются методы с эталонными алгоритмами, например, метод k-средних, то задаваемые начальные условия разбиения в значительной степени определяют конечный результат разбиения.

После завершения процедур классификации необходимо оценить полученные результаты. Для этой цели используется некоторая мера качества

классификации, которую принято называть функционалом или критерием качества. Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается экстремальное (минимальное или максимальное) значение целевой функции – функционала качества.

В большинстве случаев алгоритмы классификации и критерии качества связаны между собой, т.е. определенный алгоритм обеспечивает получение экстремального значения соответствующего функционала качества. Например, использование метода Уорда приводит к получению кластеров с минимальной внутриклассовой дисперсией.

Рассмотрим наиболее распространенные функционалы качества[13].

1. Сумма квадратов расстояний до центров классов:

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_l} d^2\left(X_i, \overline{X_l}\right),$$

где l – номер кластера (l = 1, 2,..., k), $\overline{X_l}$ – центр l-го кластера, X_i – вектор значений переменных для i-го объекта, входящего в l-й кластер, $d\left(X_i, \overline{X_l}\right)$ – расстояние между i-м объектом и центром l-го кластера.

При использовании этого критерия стремятся получить такое разбиение совокупности объектов на k кластеров, при котором значение F_1 было бы минимальным.

2. Сумма внутриклассовых расстояний между объектами:

$$F_2 = \sum_{l=1}^{k} \sum_{i \in S_l} d_{ij}^{\ 2}$$

В этом случае наилучшим следует считать такое разбиение, при котором достигается минимальное значение F_2, т.е. получены кластеры большой «плотности». Объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

3. Суммарная внутриклассовая дисперсия:

[13] Многомерный статистический анализ в экономике: Учеб. пособие для вузов / Под ред. проф. В.Н. Томашевича. – М.: ЮНИТИ-ДАНА, 1999. – С. 498–501.

$$F_3 = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{ij}^2 ,$$

где σ_{ij}^2 – дисперсия j-й переменной в кластере S_l. В данном случае разбиение, при котором сумма внутриклассовых (внутригрупповых) дисперсий будет минимальной, следует считать оптимальным. Существует несколько алгоритмов кластерного анализа, обеспечивающих оптимальное разбиение с точки зрения функционала F_3 Например, итерационный алгоритм, включающий следующие вычислительные процедуры:

а) в качестве начального разбиения задается разбиение совокупности объектов на k кластеров. Оно может быть получено одним из иерархических методов;

б) для каждого кластера S_l определяется центр $\overline{X_l} = (\overline{x_{1l}}, \overline{x_{2l}}, ..., \overline{x_{pl}})$. Каждая координата центра вычисляется следующим образом:

$$\overline{x}_{jl} = \frac{1}{n_l} \sum_{i=1}^{n_1} x_{ij}$$

где i – номер объекта, j – номер переменной, l – номер кластера, n_l – количество объектов в кластере S_l;

в) все объекты исходной совокупности распределяются по кластерам в зависимости от их расстояния до центров этих кластеров, т.е. i-й объект будет включен в кластер S_l в том случае, если его расстояние до центра этого кластера минимально:

$$d_{i\overline{x_l}} = \min_{q=1...k} \left| d_{i\overline{x_q}} \right|$$

После распределения объектов по k кластерам сравнивают первоначальный состав этих кластеров с вновь полученным. Если обнаруживается несовпадение, тогда работа алгоритма продолжается, повторяются процедуры (б) и (в). Локальный экстремум достигается в том случае, если совпадают результаты последующей и предыдущей группировок. Следует заметить, что для другого начального разбиения оптимальное значение функционала F_3 будет отличаться. На принципе

минимизации внутрикластерной дисперсии основаны алгоритмы метода k-средних и метода Уорда.

Если оценивать качество разбиения по степени удаленности кластеров друг от друга, то можно использовать функционал F_4 – средние межклассовые расстояния:

$$F_4 = \frac{\sum_{i \in S_q, j \in S_l} d_{ij}}{\sum_{l<q} n_l n_q} \cdot \max$$

Кроме названных функционалов качество классификации можно также оценивать и при помощи критерия Хотеллинга T^2 для проверки гипотезы о равенстве векторов средних для многомерных совокупностей:

$$T^2 = \frac{n_l n_q}{n_l + n_q} (\overline{X}_l - \overline{X}_q)' S_*^{-1} (\overline{X}_l - \overline{X}_q)$$

Судить о качестве разбиения позволяют и некоторые простейшие приемы. Например, сравнение средних значений признаков в отдельных кластерах (группах) со средними значениями в целом по всей совокупности объектов. Если отличие групповых средних от общего среднего значения существенное, то это может являться признаком хорошего разбиения. Оценка существенности различий может быть выполнена с помощью t-критерия Стьюдента.

Результаты многомерной классификации можно оценивать и по типу образовавшихся кластеров. Считается, что чем больше среди них кластеров типа сгущения или «сильных» кластеров, тем лучше качество разбиения.

Существуют визуальные способы исследования результатов кластеризации. Они связаны прежде всего со *свойствами кластеров*[14].

[14] Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: Учеб. пособие / Под ред. И.В. Орловой. – М.: Вузовский учебник, 2009. – С. 102–103.

Результаты классификации, полученные при использовании разных методов кластеризации, могут существенно отличаться друг от друга. При этом зависимость результатов от выбранного метода тем сильнее, чем менее явно изучаемая совокупность разделяется на «схожие группы объектов». В связи с этим целесообразно проводить классификацию по нескольким методам. Если при этом результаты, полученные по разным методам, оказываются близки, то совокупность исследуемых объектов действительно можно классифицировать. В противном случае любая классификация не является объективной.

Кластерный анализ – это не только формализуемая процедура; в нем всегда есть место наблюдению, интуиции, искусству и творчеству исследователя.