Методы кластерного анализа

Семёнова Ксения
Гр. 17.2 - 504

Таблица 13.1. Набор данных A		
№ примера	признак X	признак Y
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Данные в табличной форме не носят информативный характер. Представим переменные X и Y в виде диаграммы рассеивания, изображенной на рис. 13.1.

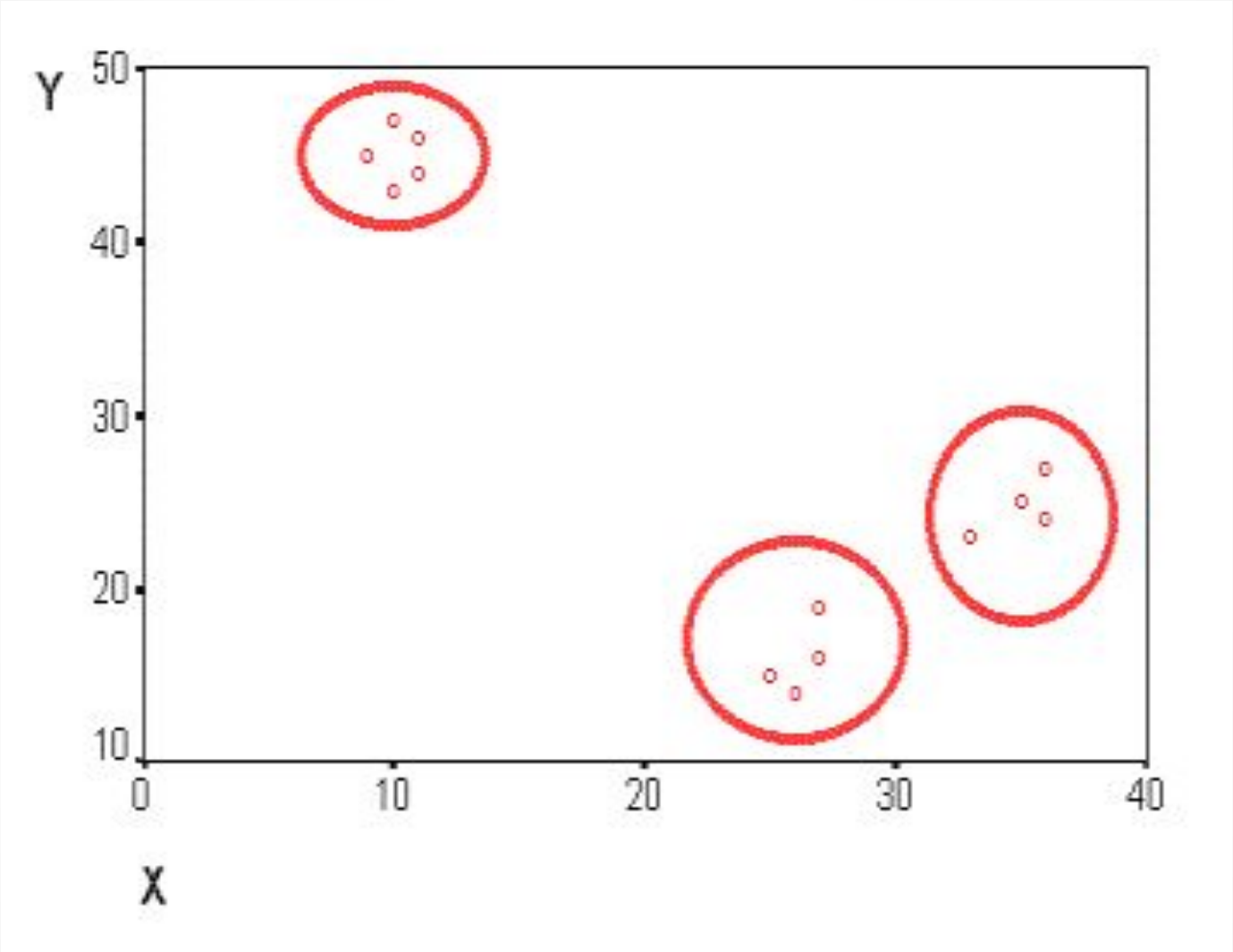


Рис. 13.1. *Диаграмма рассеивания переменных X и Y*

- *Кластер* имеет следующие ***математические характеристики***: *центр, радиус, среднек вадратическое отклонение, размер кластера*.
- ***Центр кластера*** - это среднее геометрическое *место* точек в пространстве переменных.
- ***Радиус кластера*** - максимальное *расстояние* точек от *центра кластера*.

- ***Спорный объект*** - это *объект*, который по мере сходства может быть отнесен к нескольким кластерам.
- ***Размер кластера*** может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера. *Объект* относится к кластеру, если *расстояние* от объекта до *центра кластера* меньше *радиуса кластера*. Если это условие выполняется для двух и более кластеров, *объект* является *спорным*.

Методы кластерного анализа

- иерархические;
- неиерархические.

Иерархические методы кластерного анализа

- Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

- Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.
- В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)

- Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде *дендрограммы* показан на рис. 13.3.

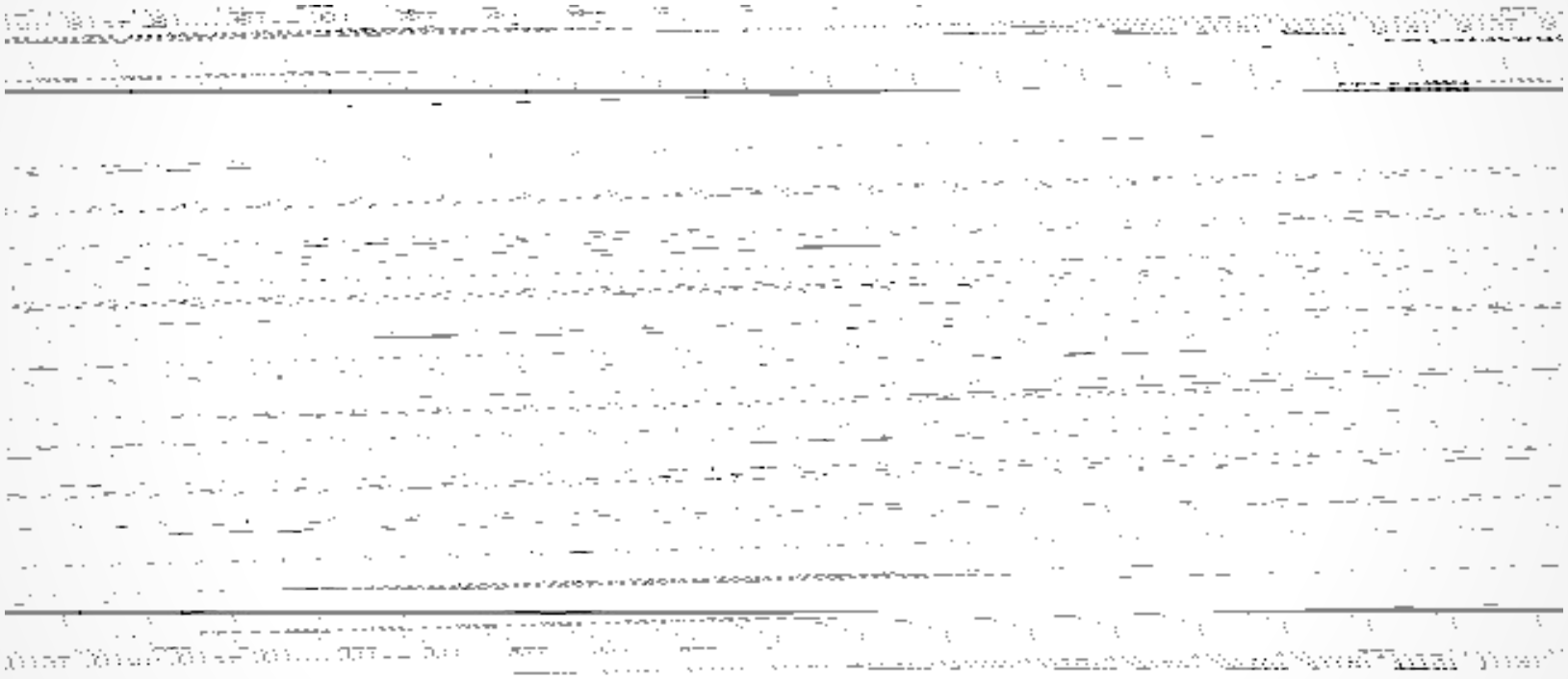
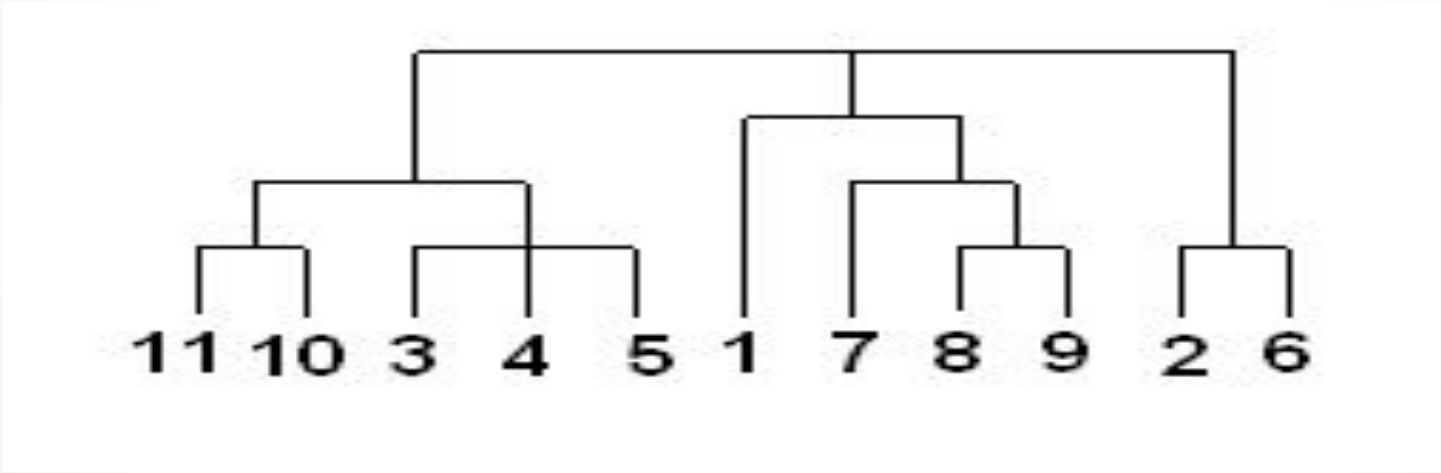


Рис. 13.3. Дендрограмма агломеративных и дивизимных методов

- Иерархические алгоритмы связаны с построением *дендрограмм* (от греческого dendron - "дерево"), которые являются результатом *иерархического кластерного анализа*.
- ***Дендрограмма*** описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.
- ***Дендрограмма*** (*dendrogram*) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Существует много способов построения *дендрограмм*. В *дендрограмме* объекты могут располагаться вертикально или горизонтально. Пример вертикальной *дендрограммы* приведен на рис. 13.4



Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Меры сходства

- Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. *Евклидово расстояние* наиболее популярная *мера* сходства.
- Квадрат *евклидова расстояния*.
- Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом *евклидова расстояния* путем возведения в квадрат стандартного *евклидова расстояния*.

Наиболее распространенный способ - *вычисление евклидова расстояния* между двумя точками i и j на плоскости, когда известны их *координаты* X и Y:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

Примечание: чтобы узнать *расстояние* между двумя точками, надо взять разницу их координат по каждой оси, возвести ее в квадрат, сложить полученные значения для всех осей и извлечь квадратный корень из суммы.

- Когда осей больше, чем две, *расстояние* рассчитывается таким образом: сумма квадратов разницы координат состоит из стольких слагаемых, сколько осей (измерений) присутствует в нашем пространстве. Например, если нам нужно найти *расстояние* между двумя точками в пространстве трех измерений (такая ситуация представлена на рис. 13.2), формула (13.1) приобретает вид:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2},$$

(13.2)

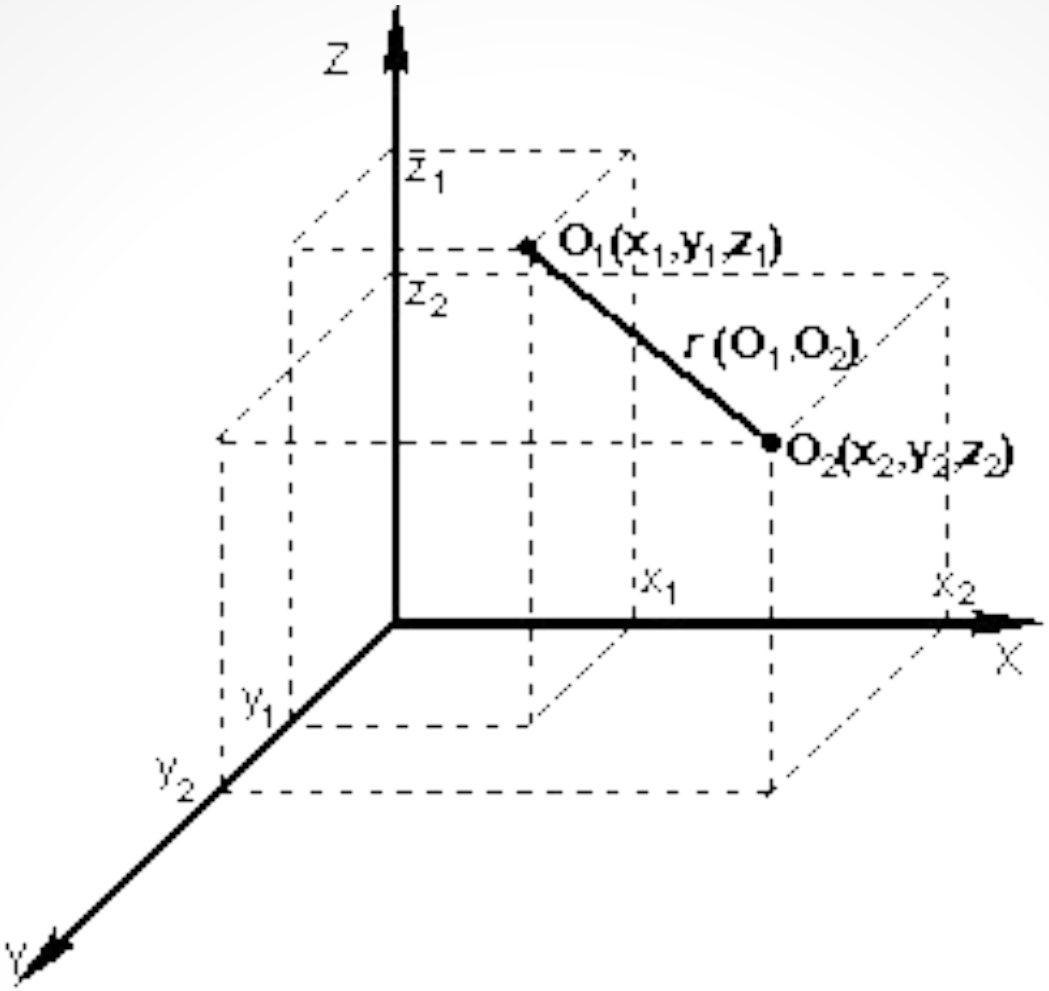


Рис. 13.2. Расстояние между двумя точками в пространстве трех измерений

Манхэттенское расстояние (*расстояние* городских кварталов),также называемое "хэмминговым" или "сити-блок" расстоянием.

- Это *расстояние* рассчитывается как среднее разностей по координатам. В большинстве случаев эта *мера* расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании *евклидова расстояния*, поскольку здесь *координаты* не возводятся в квадрат.

- **Расстояние Чебышева**.
  Это *расстояние* стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.
- **Процент несогласия**.
  Это *расстояние* вычисляется, если данные являются категориальными.

Методы объединения или связи

- Когда каждый *объект* представляет собой отдельный *кластер*, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос - как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

Метод ближнего соседа *или* одиночная связь

- Здесь *расстояние* между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В результате работы этого метода кластеры представляются длинными "цепочками" или "волокнистыми" кластерами, "сцепленными вместе" только отдельными элементами, которые случайно оказались ближе остальных друг к другу.

Метод наиболее удаленных соседей или полная *связь*

- Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты действительно происходят из различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать.

Метод Варда (Ward's *method*)

- В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до *центров кластеров*, получаемый в результате их объединения. В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы *дисперсионного анализа*. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т. е. внутригрупповой суммы квадратов. Этот метод направлен на *объединение* близко расположенных кластеров и "стремится" создавать кластеры малого размера.

Метод невзвешенного попарного среднего (метод невзвешенного попарного арифметического среднего - unweighted pair-group *method* using *arithmetic* averages, UPGMA (Sneath, Sokal, 1973)).

- В качестве расстояния между двумя кластерами берется среднее *расстояние* между всеми парами объектов в них. Этот метод следует использовать, если объекты действительно происходят из различных "рощ", в случаях присутствия кластеров "цепочного" типа, при предположении неравных размеров кластеров.

Метод взвешенного попарного среднего (метод взвешенного попарного арифметического среднего - weighted pair-group *method*using *arithmetic* averages, WPGM A (Sneath, Sokal, 1973)).

- Этот метод похож на метод невзвешенного попарного среднего, разница состоит лишь в том, что здесь в качестве весового коэффициента используется ***размер кластера*** (число объектов, содержащихся в кластере).
- Этот метод рекомендуется использовать именно при наличии предположения о кластерах разных размеров.

Невзвешенный центроидный метод (метод невзвешенного попарного центроидного усреднения - unweighted pair-group *method*using the *centroid average* (Sneath *and* Sokal, 1973)).

- В качестве расстояния между двумя кластерами в этом методе берется *расстояние* между их центрами тяжести.

Взвешенный центроидный метод (метод взвешенного попарного центроидного усреднения - weighted pair-group *method* using the *centroid average*, WPGMC (Sneath, Sokal 1973)).

- Этот метод похож на предыдущий, разница состоит в том, что для учета разницы между *размерами кластеров* (числе объектов в них), используются веса. Этот метод предпочтительно использовать в случаях, если имеются предположения относительно существенных отличий в *размерах кластеров*.

Иерархический кластерный анализ в SPSS

- Процедура *иерархического кластерного анализа* в SPSS предусматривает группировку как объектов (строк матрицы данных), так и переменных (столбцов). Можно считать, что в последнем случае роль объектов играют строки, а роль переменных - столбцы.

- В этом методе реализуется иерархический **агломеративный** *алгоритм*, смысл которого заключается в следующем:
- Перед началом кластеризации все объекты считаются отдельными кластерами, в ходе алгоритма они объединяются.
- Вначале выбирается пара ближайших кластеров, которые объединяются в один *кластер*.
- В результате количество кластеров становится равным N-1.
- Процедура повторяется, пока все классы не объединятся. На любом этапе *объединение* можно прервать, получив нужное число кластеров.
- Таким образом, результат работы алгоритма агрегирования зависит от **способов вычисления расстояния между объектами и определения близости между кластерами**.

Для определения расстояния между парой кластеров могут быть сформулированы различные подходы. С учетом этого в SPSS предусмотрены следующие методы:

- Среднее расстояние между кластерами (Between-groups linkage), устанавливается по умолчанию.
- Среднее расстояние между всеми объектами пары кластеров с учетом расстояний внутри кластеров (Within-groups linkage).
- Расстояние между ближайшими соседями - ближайшими объектами кластеров (Nearest neighbor).
- Расстояние между самыми далекими соседями (Furthest neighbor).
- Расстояние между *центрами кластеров* (*Centroid* clustering) или центроидный метод. Недостатком этого метода является то, что центр объединенного кластера вычисляется как среднее центров объединяемых кластеров, без учета их объема.
- Метод медиан - тот же центроидный метод, но центр объединенного кластера вычисляется как среднее всех объектов (*Median*clustering).
- Метод Варда.

Пример иерархического кластерного анализа

- Порядок агломерации (протокол объединения кластеров) представленных ранее данных приведен в таблице 13.2. В протоколе указаны такие позиции:
- *Stage* - стадии объединения (шаг);
- Cluster Combined - объединяемые кластеры (после объединения кластер принимает минимальный номер из номеров объединяемых кластеров);
- *Coefficients* - коэффициенты.

Таблица 13.2. Порядок агломерации

Cluster Combined		Coefficients
Cluster 1	Cluster 2	
9	10	,000
2	14	1,461E-02
3	9	1,461E-02
5	8	1,461E-02
6	7	1,461E-02
3	13	3,490E-02
2	11	3,651E-02
4	5	4,144E-02
2	6	5,118E-02
4	12	,105
1	3	,120
1	4	1,217
1	2	7,516

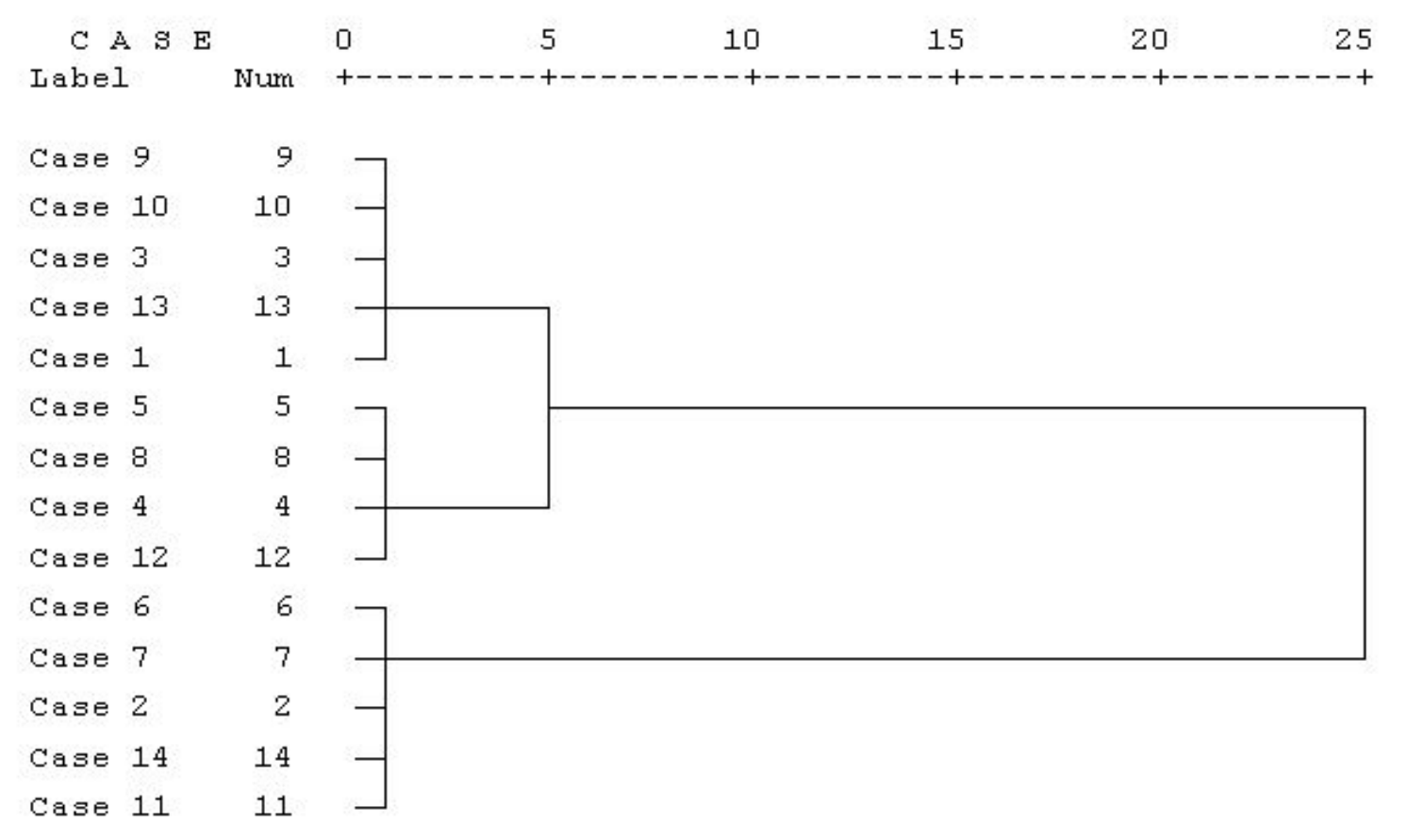
Процедура *стандартизации* используется для исключения вероятности того, что классификацию будут определять переменные, имеющие наибольший разброс значений. В SPSS применяются следующие виды *стандартизации*:

- Z-шкалы (Z-Scores). Из значений переменных вычитается их среднее, и эти значения делятся на стандартное отклонение.
- Разброс от -1 до 1. Линейным преобразованием переменных добиваются разброса значений от -1 до 1.
- Разброс от 0 до 1. Линейным преобразованием переменных добиваются разброса значений от 0 до 1.
- Максимум 1. Значения переменных делятся на их максимум.
- Среднее 1. Значения переменных делятся на их среднее.
- Стандартное отклонение 1. Значения переменных делятся на стандартное отклонение.

Определение количества кластеров

- Способ сводится к определению скачкообразного увеличения некоторого коэффициента, который характеризует переход от сильно связанного к слабо связанному состоянию объектов.

На верхней линии по горизонтали отмечены номера шагов алгоритма, всего алгоритму потребовалось 25 шагов для объединения всех объектов в один кластер.




Спасибо
за
внимание