Кластерный анализ

Минск 2003

Литература

1. Факторный, дискриминантный и кластерный анализ: Пер. с англ. / Дж.-О.Ким, Ч.У.Мюллер, У. Р.Клекка и др.; Под ред. И.С.Енюкова. – М.: Финансы и статистика, 1989.
2. Бююль А., Цефель П. SPSS: искусство обработки информации. – СПб.: ДиаСофт, 2001.

Предназначение кластерного анализа

- Построение эмпирической классификации

  Кластерный анализ позволяет разбить выборку на группы схожих объектов, называемых *кластерами*. Члены одной группы (кластера) должны обладать схожими проявлениями переменных, а члены разных групп – различными.

Стратегия кластерного анализа

1. Выбор переменных и их предварительные преобразования
2. Определение меры сходства (подобия) между объектами
3. Выбор метода кластеризации
4. Определение числа кластеров и обоснование кластерного решения

Важнейшие семейства кластерных методов

1. Иерархические методы
   1.1. Агломеративные *
   1.2. Дивизимные
2. Итеративные методы
   2.1. Метод К средних *
   2.2. Метод "восхождения на холм"
3. Факторные методы *
4. Нейросетевые методы

* Реализованы в SPSS

Пример: **классификация людей по скорости реакции на свет и звук**

1. Выбор переменных

- время реакции на свет
- время реакции на звук

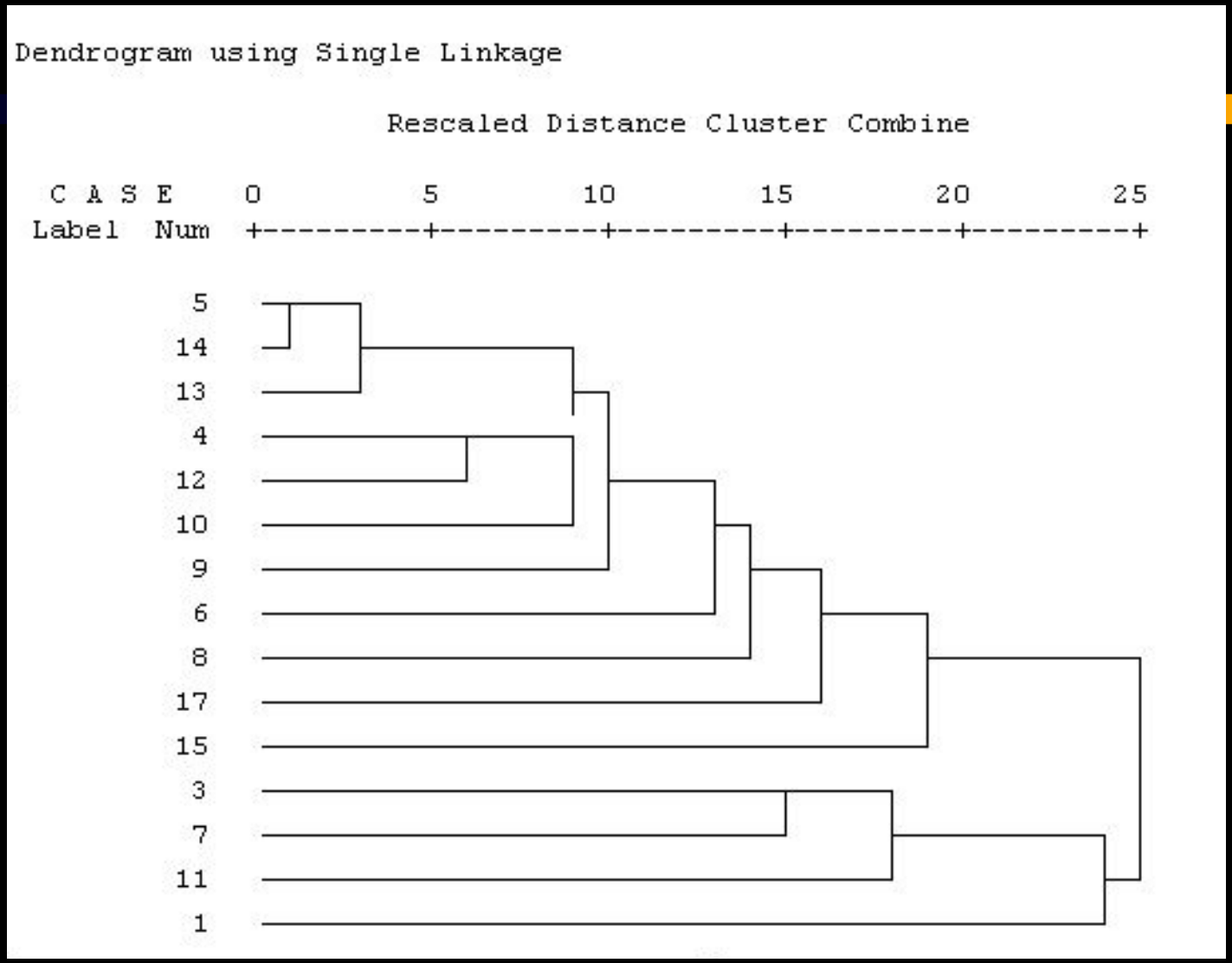
2. Выбор меры сходства

- Евклидово расстояние

3. Выбор метода кластеризации

- агломеративный
  - метод одиночной связи
    (метод "ближайшего соседа")

Дендрограмма



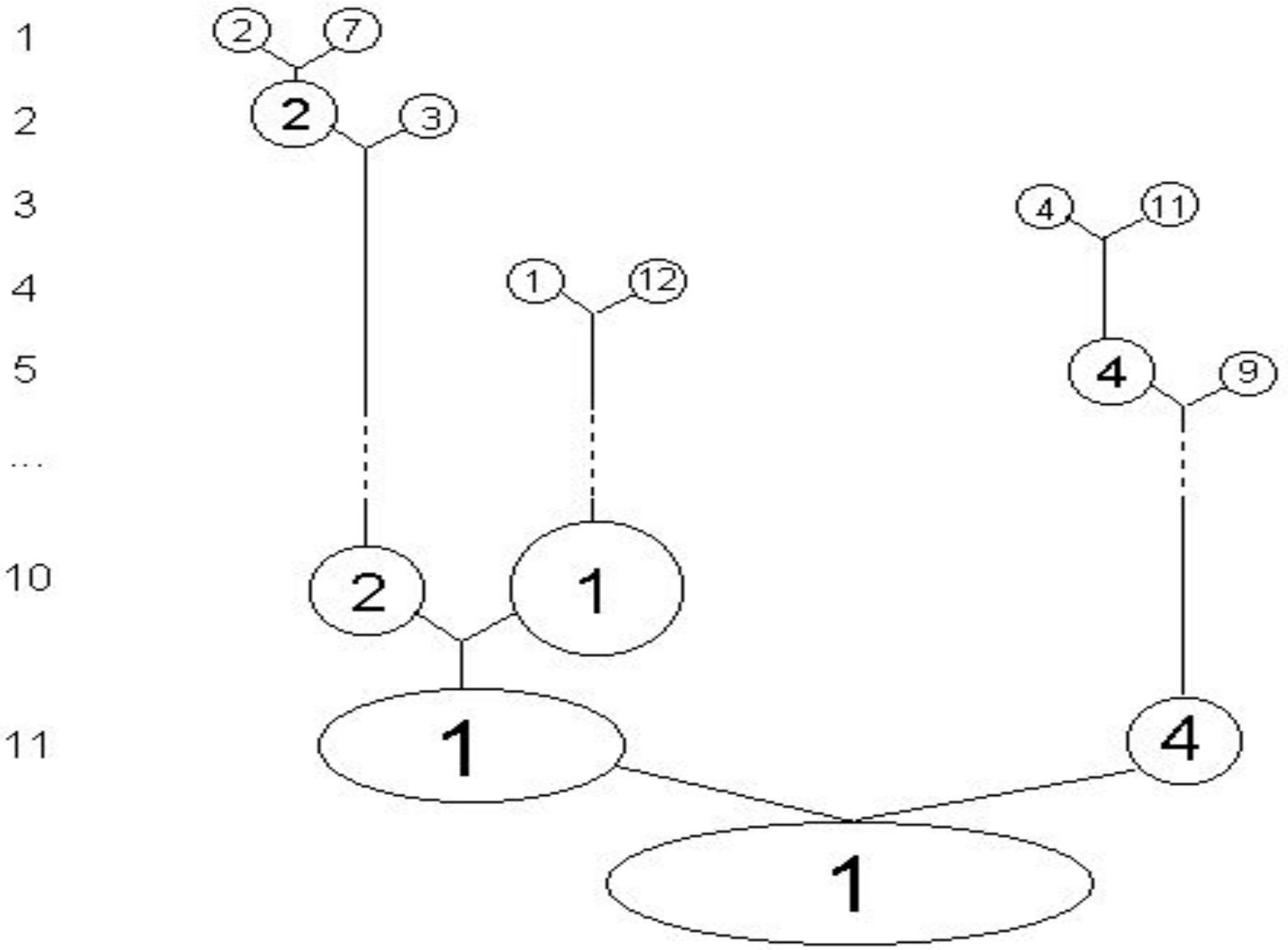
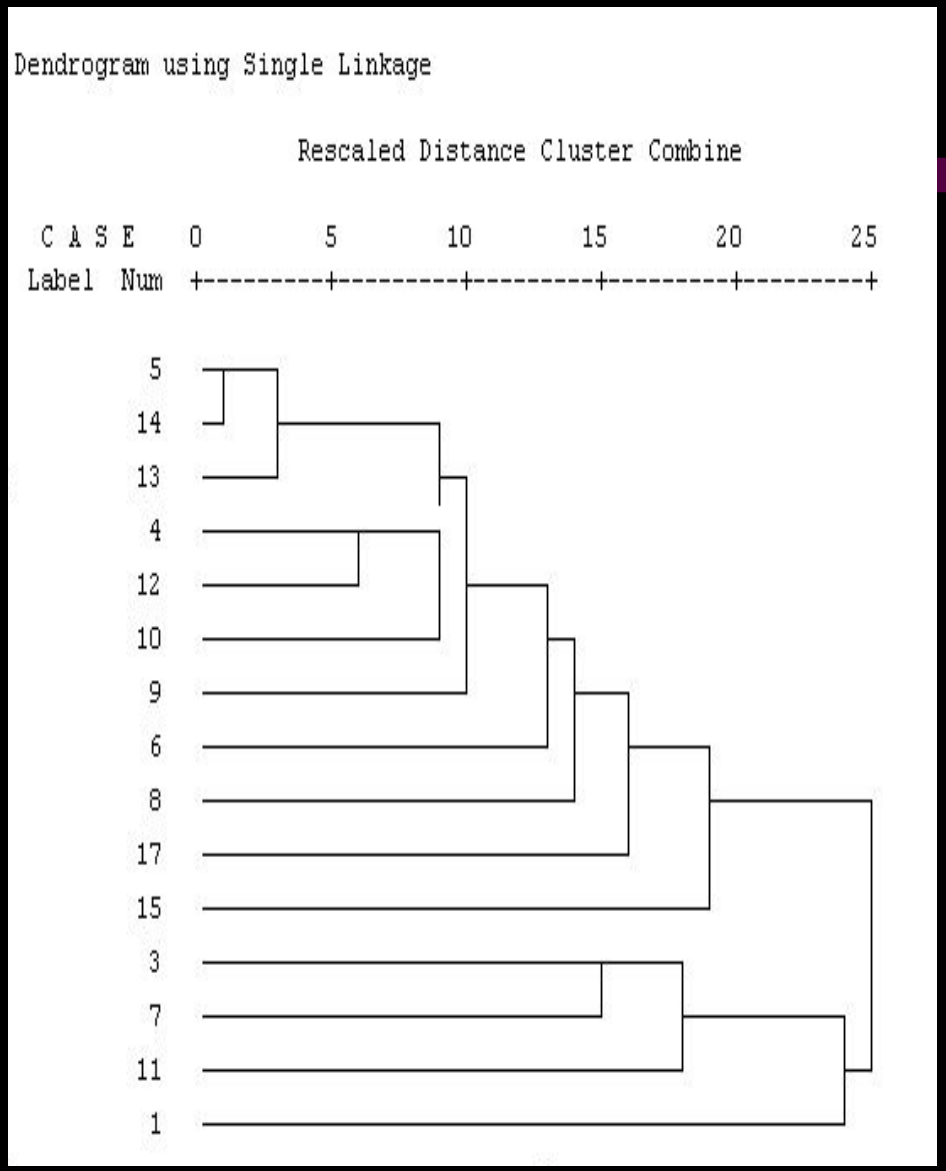
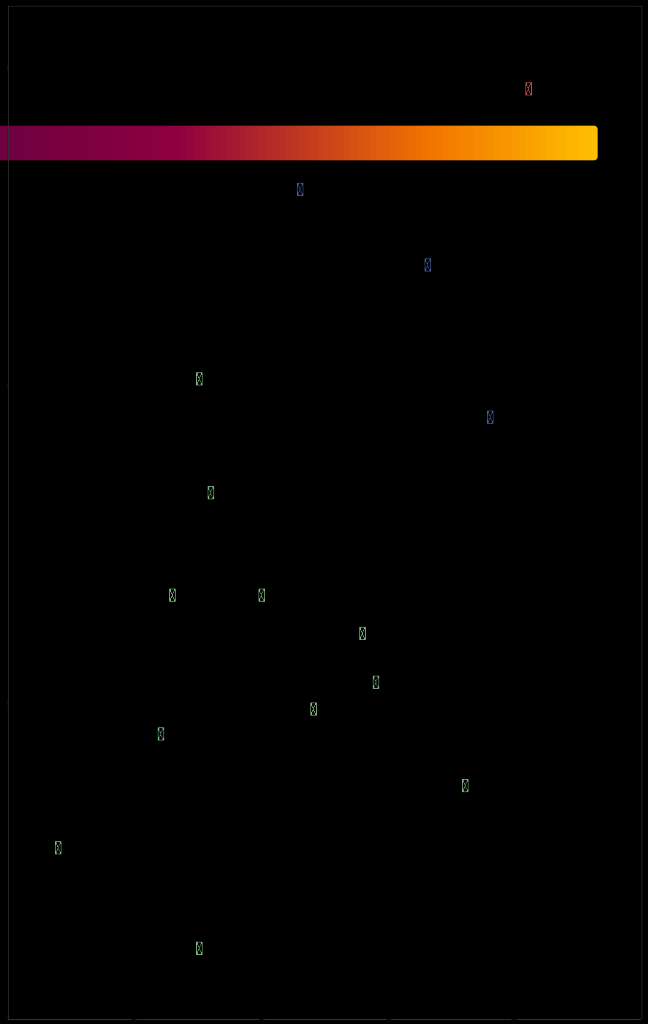
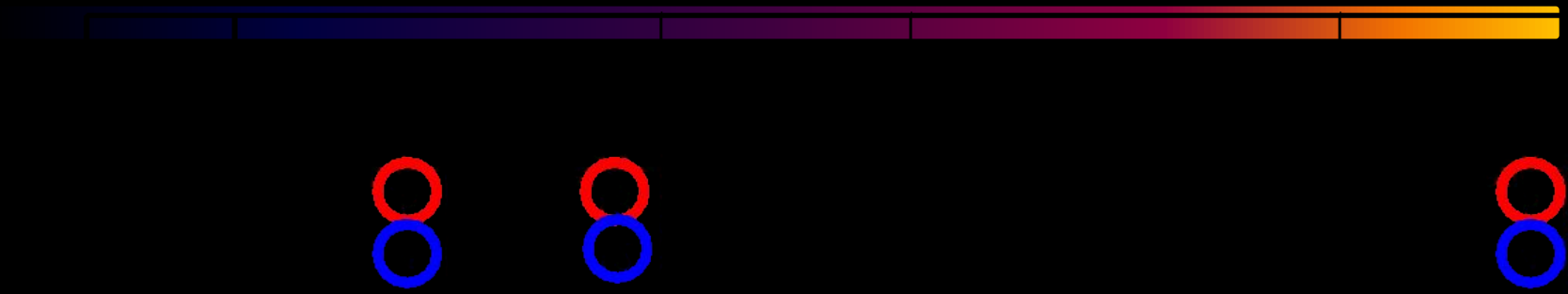

1
2
3
4
5
...
10
11
2
7
2
3
1
12
4
11
4
9
2
1
1
4
1

Диаграмма рассеяния





Порядок объединения



Сколько кластеров оставить

- Там где расстояние между кластерами (колонка «coefficients»), определенное на основании выбранной меры увеличивается скачкообразно, процесс необходимо остановить, так как будут объеденены кластеры находящиеся слишком далеко друг от друга.

Методы объединения или связи

- На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако когда **связываются вместе несколько объектов**, возникает вопрос, как следует определить расстояния между кластерами?
- Другими словами, необходимо правило объединения или связи для двух кластеров. Здесь имеются различные возможности.

Методы объединения кластеров

1. Метод одиночной связи
2. Метод полной связи
3. Метод средней связи
   - 3.1. Невзвешенный
   - 3.2. Взвешенный (внутригрупповой)
4. Центроидный метод
   - 4.1. Невзвешенный
   - 4.2. Взвешенный (медианый)
5. Метод Уорда

Меры подобия в SPSS

- Евклидова дистанция. Наиболее простой и легко интерпретируемый путь. Следует помнить, что на расстояния могут сильно влиять различия между осями, по координатам которых вычисляются эти расстояния. К примеру, если одна из осей измерена в сантиметрах, а вы потом переведете ее в миллиметры (умножая значения на 10), то окончательное евклидово расстояние (или квадрат евклидова расстояния), вычисляемое по координатам, сильно изменится, и, как следствие, результаты кластерного анализа могут сильно отличаться от предыдущих. При этом следует использовать z-стандартизацию.

- **Квадрат евклидова расстояния.** Иногда стандартное евклидово расстояние возводят в квадрат, чтобы придать большие веса более отдаленным друг от друга объектам.
- **Расстояние Чебышева.** Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (т.е. каким-либо одним признаком).
- **Процент несогласия.** Эта мера используется в тех случаях, когда данные являются категориальными.