Министерство образования и науки Российской Федерации
ФГАОУ ВО «Севастопольский государственный университет»

Институт информационных технологий
и управления в технических системах

Лабораторная работа №4
«Кластерный анализ. Основные этапы и задачи кластерного анализа данных»

по дисциплине «Интеллектуальный анализ данных»
для студентов всех форм обучения направления подготовки
09.03.02 «Информационные системы и технологии»

Севастополь
2017

Кластерный анализ. Основные этапы и задачи кластерного анализа данных. Методические указания к лабораторным занятиям по дисциплине «Интеллектуальный анализ данных» / Сост.: О.А. Сырых – Севастополь: Изд-во СевГУ, 2017 – 22 с.

Методические указания предназначены для проведения лабораторных работ по дисциплине «Интеллектуальный анализ данных». Целью методических указаний является помощь студентам в изучении возможностей системы RStudio. Излагаются практические сведения необходимые для выполнения лабораторной работы, требования к содержанию отчета.

Методические указания рассмотрены и утверждены на методическом семинаре и заседании кафедры «Информационные системы» (протокол № 1 от 29 августа 2016 г.)

Лабораторная работа №4
Кластерный анализ. Основные этапы и задачи кластерного анализа данных.

Цель:

- Закрепить теоретические знания и приобрести практические навыки в проведении кластерного анализа по экспериментальным данным
- исследовать возможности языка R для проведения кластерного анализа.

Время: 6 часов

Лабораторное оборудование: персональные компьютеры, выход в сеть Internet, RStudio.

Краткие теоретические сведения

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

Цель кластеризации - поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Само понятие "кластер" определено неоднозначно: в каждом исследовании свои "кластеры". Переводится понятие кластер (cluster) как "скопление", "гроздь".

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Термин кластерный анализ, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов.

Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной.

Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике.

Методов вычисления расстояний существует очень много (не забывайте, что дело происходит в многомерном пространстве). Наиболее широко употребляемыми методами для непрерывных переменных являются: эвклидово расстояние (рис. 1) – Euclidian distances (1) и манхеттеновское расстояние или расстояние городских кварталов – City-block (Manhattan) (2).

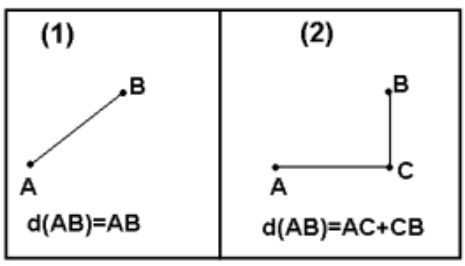

Рисунок – 1. Методов вычисления расстояний.

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

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

Методы кластерного анализа можно разделить на две группы:

– иерархические;

– неиерархические.

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. 2.

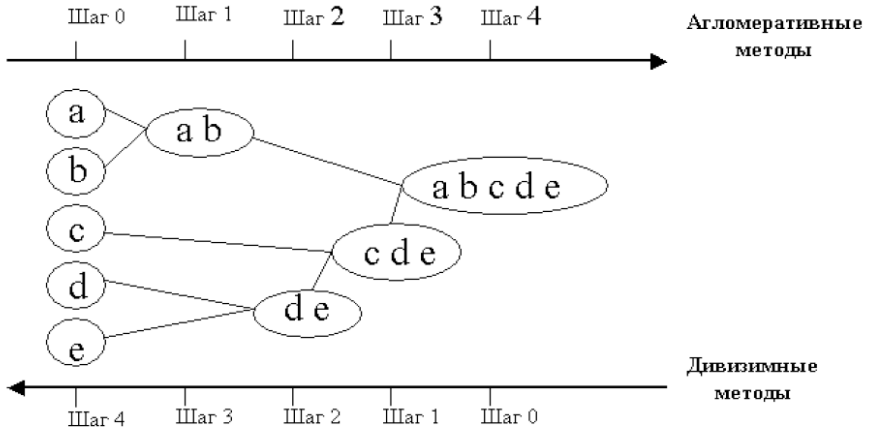

Рисунок – 2. Дендрограмма агломеративных и дивизимных методов.

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

Преимуществом иерархических методов кластеризации является их наглядность.

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.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Итеративные методы

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов.

Алгоритм k-средних (k-means)

Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом.

В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Описание алгоритма

1. Первоначальное распределение объектов по кластерам.

Выбирается число *k*, и на первом шаге эти точки считаются "центрами" кластеров.

Каждому кластеру соответствует один центр.

Выбор начальных центроидов может осуществляться следующим образом:

– выбор k-наблюдений для максимизации начального расстояния;
– случайный выбор k-наблюдений;
– выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

2. Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

– кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
– число итераций равно максимальному числу итераций.

На рис. 3 приведен пример работы алгоритма k-средних для k, равного двум.

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга).

Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Достоинства алгоритма k-средних:

– простота использования;
– быстрота использования;
– понятность и прозрачность алгоритма.

Недостатки алгоритма k-средних:

– алгоритм слишком чувствителен к выбросам, которые могут искажать среднее.

Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;

– алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

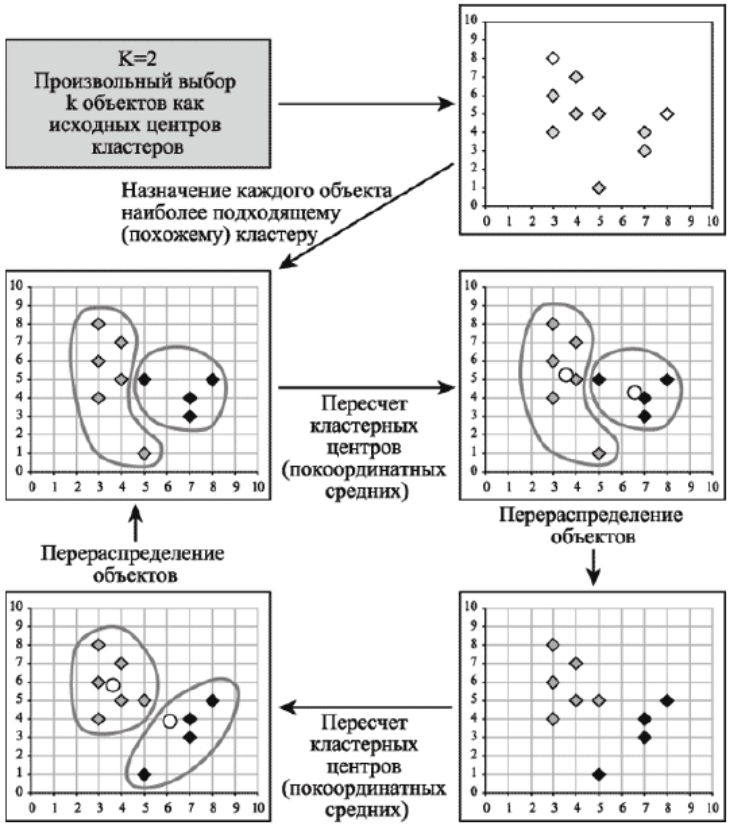

Рисунок - 3. Пример работы алгоритма k-средних (k=2)

Задание и порядок выполнения лабораторной работы №4

Кластерный анализ методом k-средних в R

1. Создать файл с исходными данными.
2. Кластерный анализ проводится в пакете Rcmdr

   Статистика

   Многомерный анализ

   Кластерный анализ

   Кластерный анализ k-средних. В опциях есть возможность выбрать количество кластеров.

   Функция кластерного анализа в R:

```
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))
```

Аргументы:

x – численная матрица, содержащая объекты;

centers – или число кластеров, или множество исходных центров кластеров. Если аргумент представляет собой число, то выбирается случайное множество центро кластеров;

iter.шах – максимальное число итераций;

nstart – если centers – число, то данный аргумент определяет, как много случайных множеств может быть выбрано;

algorithm – символ, определяющий используемый алгоритм.

Возвращаемое значение: Объект класса kmeans, который представляет собой список следующих компонент:

cluster – вектор целых чисел, определяющих, в каком кластере размещены объекты;

centers – матрица центров кластеров;

withinss – сумма квадратов расстояний между точками для каждого кластера;

size – число точек в каждом кластере.

3. Провести кластерный анализ экспериментальных данных.

4. Проведя процедуру кластеризации (разбиение на классы или кластеры) несколько раз при различных значениях числа кластеров (от 2-х до 10 кластеров), необходимо выбрать лучшую группировку в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний:

$$F = \frac{d_w / f_w}{d_b / f_b}.$$

Для сравнения нескольких типизаций и выбора наиболее оптимальной из них необходим критерий, численная мера качества классификации.

Одна из оценок качества служит показатель

$$J = J_1 \big/ J_2,$$

где

$$J_1 = \frac{2}{m(m-1)} \sum_{i=1}^{m} \overline{D}_{ii} \quad J_2 = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \overline{D}_{ij} \quad .$$

Здесь $\overline{D}_{ii}$ – среднее расстояние между точками внутри i-го класса, $\overline{D}_{ij}$ – среднее расстояние между парами точек i-го и j-го классов, где m – количество кластеров разбиения.

Полученные результаты оформите в виде Таблицы.

Изобразinm графически значения данного показателя качества классификации. Для этого построить диаграмму, на которой по оси X – количество кластеров, по оси Y – значения показателя J.

Для графической интерпретации используем критерий "каменистой осыпи". Обычно, для выбора размерности какого-либо пространства, используют график зависимости стресса от размерности (график каменистой осыпи). Этот критерий впервые был предложен Кэттелом (Cattell (1966)) в контексте решения задачи снижения размерности в факторном анализе.

Кэттел предложил найти такую абсциссу на графике, в которой график стресса начинает визуально сглаживаться в направлении правой, пологой его части, и, таким образом, уменьшение стресса максимально замедляется. Образно говоря, линия на рисунке напоминает скалистый обрыв, а черные точки на графике напоминают камни, которые ранее упали вниз. Таким образом, внизу наблюдается как бы каменистая осыпь из таких точек. Справа от выбранной точки на оси абсцисс, лежит только "факторная осыпь".

5. Сформулировать выводы.

Пример:

1. Использован файл данных Данные.xls
2. Проведено разбиение на 2 кластера (рис 4)

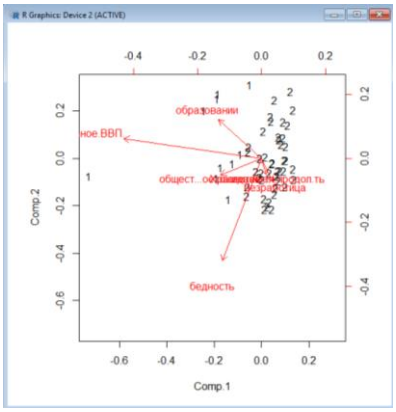

Рисунок 4 – Разбиение данных на 2 кластера

Результаты выполненного анализа:

```
> .cluster$size # Cluster Sizes
[1] 10 44

> .cluster$centers # Cluster Centroids
  new.x.Х.ожидаемая.продол.ть new.x.бедность new.x.безработица new.x.образовании
1                    0.6820000      0.7430000         0.8430000          0.9100000
2                    0.6972727      0.7286364         0.8688636          0.8390909
  new.x.общест...ое.развитие new.x.реальное.ВВП.
1                  0.6862000           0.2528000
2                  0.6493182           0.1129318

> .cluster$withinss # Within Cluster Sum of Squares
[1] 0.2077872 0.2390601

> .cluster$tot.withinss # Total Within Sum of Squares
[1] 0.4468473

> .cluster$betweenss # Between Cluster Sum of Squares
[1] 0.2204887
```

– количество элементов в кластере: первый кластер содержит 10 элементов, второй – 44;
– сумма квадратов расстояний внутри кластера: 1 – 0,208, 2 – 0,239
– общая сумма квадратов расстояний внутри кластеров: 0,447
– сумма квадратов расстояний между кластерами – 0,22.

3. Для выбора лучшей группировки в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний было проведено деление на 3 – 10 кластеров и заполнена таблица в MS Excel.

Таблица.

Расчет численного показателя мера качества классификации

Кластеры	D_{ii}	m	D_{ij}	J_1	J_2	J
2	44,01	2	1,43	44,01	0,72	61,47
3	40,03	3	3,83	13,34	1,28	10,47
4	38,64	4	8,62	9,66	2,16	4,48
5	36,36	5	12,07	7,27	2,41	3,01
6	33,21	6	21,03	5,54	3,50	1,58
7	30,43	7	30,84	4,35	4,41	0,99
8	28,66	8	42,84	3,58	5,36	0,67
9	28,11	9	51,65	3,12	5,74	0,54
10	25,41	10	74,03	2,54	7,40	0,34

Графически значения данного показателя качества классификации представлено графически на рис 5. Для этого построена диаграмма, на которой по оси *X* – количество кластеров, по оси *Y* – значения показателя *J*.

Рисунок – 5. Диаграмма численной меры качества классификации

В соответствии с этим критерием оптимальным разбиением экспериментальных данных является разбиение на 3 кластера.

Иерархический кластерный анализ

1. Провести иерархический кластерный анализ в среде Rcmdr
   Статистика
   Многомерный анализ
   Кластерный анализ
   Иерархический кластерный анализ
2. Провести анализ экспериментальных данных используя разные методы. Полученные результаты сравнить и сделать выводы.
3.

Контрольные вопросы

1. Кластерный анализ.
2. Виды кластерного анализа.
3. Принцип иерархического кластерного анализа
4. Сущность метода К средних?
5. Какие существуют меры расстояний между объектами при древовидной кластеризации?

Библиография

1. Алексей Шипунов и др. Наглядная статистика. Используем R! – М.: ДМК Пресс, 2014. – 298 с. [Электронный ресурс]. Режим доступа: http://ashipunov.info/shipunov/school/books/rbook.pdf.

2. Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Издательство Российского университета дружбы народов, 2010. – 207 с.
3. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.
4. Оффициальный сайт RStudio. Режим доступа: https://www.rstudio.com.
5. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. Режим доступа: http://machinelearning.ru.
6. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга. Режим доступа: http://r-analytics.blogspot.com