Практическая работа №5: Исследование методов кластер-анализа

Цель работы

Ознакомиться с методами кластер-анализа на основе языка R.

Основные теоретические положения

Термин кластерный анализ (впервые понятие введено математиком Р. Трионом, 1939) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры, т.е. развернуть таксономии. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т.е. с обезьянами), чем с «отдаленными» членами семейства млекопитающих (например, собаками) и т.д.

Фактически, кластерный анализ является не столько обычным статистическим методом, сколько «набором» различных алгоритмов «распределения объектов по кластерам». Существует точка зрения, что в отличие от многих других статистических процедур, методы кластерного анализа используются в большинстве случаев тогда, когда вы не имеете каких-либо априорных гипотез относительно классов, но все еще находитесь в описательной стадии исследования. Следует понимать, что кластерный анализ определяет «наиболее возможно значимое решение».

Деревья кластеризации. Назначение алгоритма построения деревьев кластеризации заключается в постепенном объединении объектов в достаточно большие кластеры, используя меры расстояния и сходства между объектами. На первом шаге каждый объект является кластером. При переходе к следующему шагу группы объектов объединяются в кластеры на основе меры расстояния и выбранного метода. На каждом следующем шаге процедура повторяется для наиболее «близких» друг к другу кластеров. Использующиеся меры расстояния между объектами:

- Евклидово расстояние: $d(x,\,y) = \|\mathbf x + \mathbf y\| = \sqrt{\sum\limits_i(x_i - y_i)^2}$.
- Манхэттенское расстояние (расстояние городских кварталов): $d(x,\,y) = \sum\limits_i|x_i - y_i|$.
- Расстояние Чебышева: $d(x,\,y) = \max|x_i - y_i|$.

Использующиеся способы объединения кластеров:

- Одиночная связь (метод ближайшего соседа). В этом методе расстояние между двумя

кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах;
- Полная связь (метод наиболее удаленных соседей). В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах, то есть «наиболее удаленными соседями».
- Невзвешенное попарное среднее. В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.
- Метод Варда. Метод минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге.

Метод k-means (k-средних). Предположим, уже имеются гипотезы относительно числа кластеров (по наблюдениям или по переменным). Можно указать системе образовать ровно три кластера так, чтобы они были настолько различны, насколько это возможно. Это именно тот тип задач, которые решает алгоритм метода *k*-means. В общем случае метод *k*-means строит ровно *k* различных кластеров, расположенных на возможно больших расстояниях друг от друга.

Общая формулировка задачи

- Провести исследование на имеющихся данных, а также на модельных примерах.
- Исследовать методы иерархической группировки.
- Провести исследование иерархической кластеризации при использовании кластеризации по признакам.

Порядок выполнения работы

1. Исследование методов иерархической группировки

1. Выбрать переменные (переменная-номер класса не участвует в обработке).
2. Установить параметр Custer = CASES.
3. Установить параметр Input = RAW DATA.
4. Выбрать метод связывания Amalgamation (linkage) rule.
5. Выбрать метод измерения расстояния - Distance measure.
6. Запустить процедуру кластеризации.
7. Просмотреть результаты построения иерархического дерева. Проанализировать какие данные образуют кластеры (опция - Amalgamation Shedule).
8. Исследовать процесс кластеризации при различных сочетаниях методов связывания и методов измерения расстояния.

2. Исследование иерархической кластеризации при использовании кластеризации по признакам

1. Выбрать в меню CLUSTER ANALISYS в опции CLUSTER значение Variables= COLUMNS.
2. Провести кластеризацию аналогично п.п. 4-7 предыдущего пункта.
3. Определить наборы признаков наиболее и наименее связанные друг с другом.
4. Выбрать наиболее информативный минимальный набор признаков и проверить его

эффективность в режиме Discriminant Analysis.

3. Метод K-средних

1. Выбрать метод кластеризации K-Means Clustering (метод K-средних)
2. Загрузить исходные данные.
3. Выбрать переменные.
4. Выполнить кластер-анализ.
5. Проанализировать результаты кластеризации (K-means Clustering Results).
    - анализ дисперсии (Analysis of variance);
    - математические ожидания и евклидовы расстояния между кластерами (cluster means & Euclidian distance);
    - графики математических ожиданий по кластерам;
    - дискриптивные статистики по кластерам;
    - содержание кластеров (members of each cluster & distance).
6. Провести кластеризацию методом K-средних по признакам.
7. Сохранить графики и таблицы результатов исследования.
8. Сделать сравнительные выводы по проведенным исследованиям.
9. Оформить результаты в виде отчета.