# РГЗ. Кластеризация

Цель работы: изучение методов кластеризации.

Среда выполнения: SPSS Statistics, Statistica, Deductor, RStudio.

Задание

1. Выбрать массив данных (рекомендуются базы по ссылкам № 3-4 из списка литературы), описать параметры. Примечание: тип задачи «классификация», «кластеризация».
2. Провести дескриптивный анализ, оценить близость выборок к нормальной. В соответствии с результатами обосновать выбор методов кластеризации.
3. Стандартизировать переменные.
4. Изучить пример решения задачи кластеризации (ссылка №2 из списка литературы).
5. Построить диаграммы рассеивания (составные диаграммы рассеивания, категоризованные диаграммы рассеяния) по выбранным переменным. Интерпретировать результаты, оценить возможное количество кластеров.
6. Решить задачу кластеризации двумя методами (*k*-средних, иерархический, EM, DBSCAN, карта Кохонена и др).
7. Оценить качество построенных моделей (в т.ч. расстояние между кластерами, внутрикластерные расстояния, компактность кластеров, центры кластеров и т.д.).
8. Провести сравнительный анализ решений.
9. Исследовать влияние параметров одного из методов на качество решения, оценить полученные результаты.
10. Интерпретировать результаты.
11. Оформить отчет.

Содержание отчета

1. Титульный лист.
2. Цель работы.
3. Описание исходных данных.
4. Результаты дескриптивного анализа.
5. Диаграммы рассеивания.
6. Интерпретация результатов (количество кластеров).
7. Обоснование выбора методов кластеризации.
8. Параметры выбранных методов.
9. Оценка адекватности полученных решений.
10. Сравнительный анализ решений и интерпретация результатов.
11. Результаты исследования влияния параметров алгоритма на качество решения.

Список литературы и ссылки на материалы


1. Айвазян С.А. Методы эконометрики. – М.: Магистр: - ИНФРА-М, 2010. – 512 с.
2. Пример решения задачи кластеризации в Statistica. http://statsoft.ru/solutions/ExamplesBase/branches/detail.php?ELEMENT_ID=1573
3. Массивы данных. https://www.kdnuggets.com/datasets/index.html
4. Массивы данных. http://archive.ics.uci.edu/ml/datasets.html
5. Обзор методов кластеризации. https://habrahabr.ru/post/101338/
6. Категоризованные графики. http://statsoft.ru/home/textbook/modules/stgraph.html#categorized4
7. Методы кластеризации. http://www.machinelearning.ru/wiki/images/archive/2/28/20150427184336%21Voron-ML-Clustering-slides.pdf
8. Обучение без учителя. https://habrahabr.ru/company/ods/blog/325654/
9. Видео-курс. Кластеризация. https://ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie/lecture/o4Ij7/klastierizatsiia


Вопросы к защите

1. Методы кластеризации. Метод k-средних. Иерархические методы. Условия применения методов.
2. Меры оценки расстояния, близости объектов и кластеров.
3. Оценка качества кластеризации.