Санкт-Петербургский Политехнический Университет Петра Великого

Институт прикладной математики и механики

Кафедра Телематики при ЦНИИ РТК

Лабораторная работа

по теме

«Кластеризация»

Преподаватель Уткин Л.В.

Студент гр.43607/2 Лисенкова А.А

Санкт-Петербург

2018 г.

# Постановка задачи

В ходе выполнения лабораторной работы необходимо выполнить следующие задачи:

1. Разбить множество объектов из набора данных pluton в пакете «cluster» на 3 кластера методом центров тяжести (kmeans). Сравнить качество разбиения в зависимости от максимального числа итераций алгоритма.
2. Сгенерировать набор данных в двумерном пространстве, состоящий из 3 кластеров, каждый из которых сильно “вытянут” вдоль одной из осей. Исследовать качество кластеризации методом clara в зависимости от 1) использования стандартизации; 2) типа метрики. Объясните полученные результаты.
3. Построить дендрограмму для набора данных votes.repub в пакете «cluster» (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31).
4. Построить дендрограмму для набора данных animals в пакете «cluster». Данные содержат 6 двоичных признаков для 20 животных. Переменные - [ , 1] war теплокровные; [ , 2] fly летающие; [ , 3] ver позвоночные; [ , 4] end вымирающие; [ , 5] gro живущие в группе; [ , 6] hai имеющие волосяной покров.
5. Рассмотреть данные из файла seeds\_dataset.txt, который содержит описание зерен трех сортов пшеницы: Kama, Rosa and Canadian. Признаки: 1. область A, 2. периметр P, 3. компактность C = 4\*pi\*A/P^2, 4. длина зерна, 5. ширина зерна, 6. коэффициент ассиметрии, 7. длина колоска.

# Ход работы

**1 задание**

Из исходного множества объектов выбрано 2 признака: Pu241 и Pu240. После чего, на основе метода k средних полученные данные разбиты на 3 кластера. В зависимости от максимального числа итераций построен график, приведённый на Рис. 1. Объекты на графике представлены в 3 состояниях: при iter.max = {1,1000,2000} соответственно размеру. Можно заметить, что один объект был отнесен к разным кластерам при различных значениях iter.max .

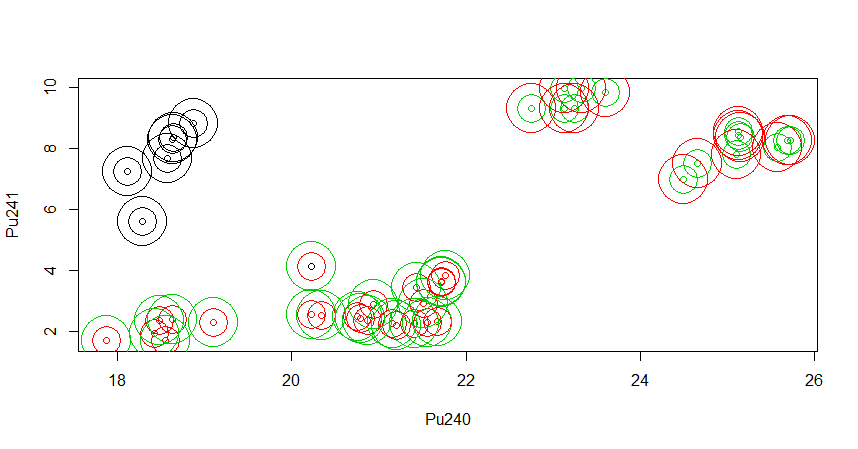


Рис. 1. Разбиение множества объектов на кластеры для примера «Pluton»

**2 задание**

Для выполнения задания сгенерирован набор данных, состоящий из трех кластеров, каждый из которых сильно вытянут вдоль одной из осей. Методом clara произведена кластеризация данных с использованием 2 типов метрик: manhattan, euclidean, - а также с использованием стандартизации и без. Полученные результаты приведена на Рис.2 .

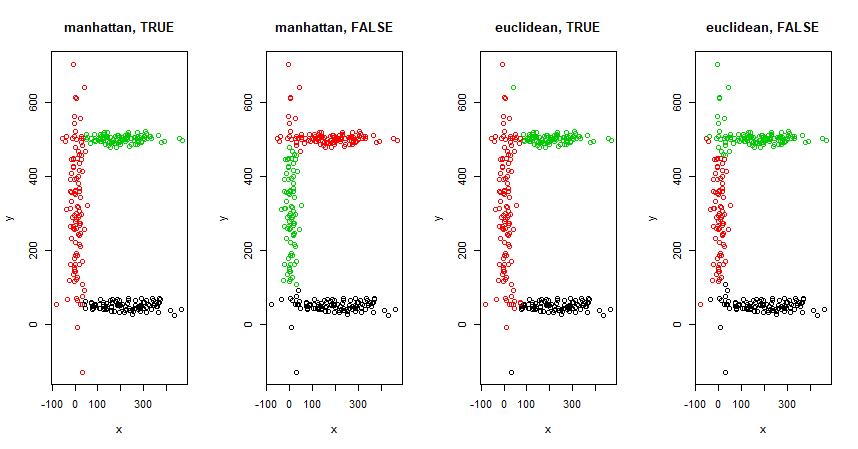


Рис. 2.Разбиение множества объектов на кластеры для примера «Points»

**3 задание**

Для указанного набора данных votes построена дендограмма (см. Рис.3). Она показывает степень близости отдельных объектов и кластеров, а также наглядно демонстрирует в графическом виде последовательность их объединения или разделения.

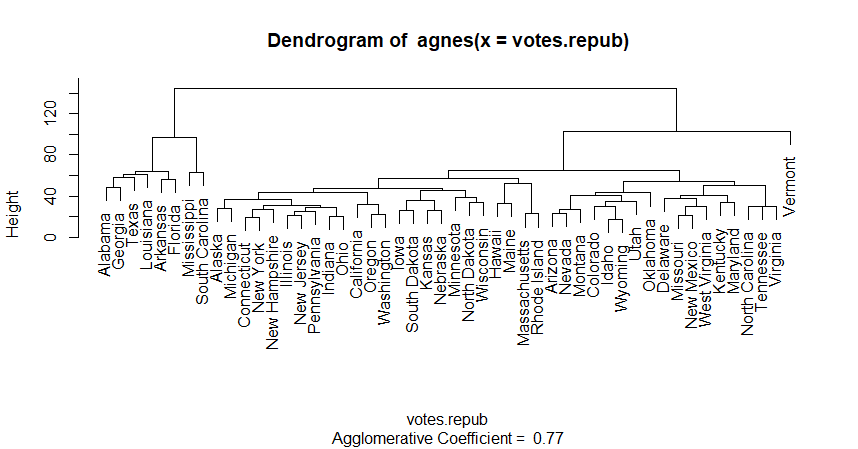


Рис. 3. Дендограмма для примера «Votes»

**4 задание**

Для заданного набора данных animals построена дендограмма (см. Рис.4). Количество уровней её соответствует числу шагов слияния или разделения кластеров.

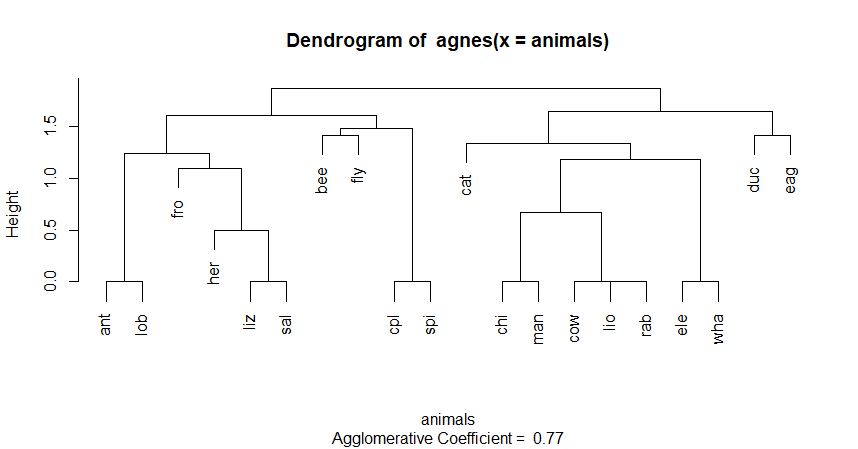


Рис. 4. Дендограмма для примера «Animals»

**5 задание**

После рассмотрения данных, отображающих описание зерен трёх сортов пшеницы, построено соответствующие разбиение на 3 кластера (см. Рис.5).

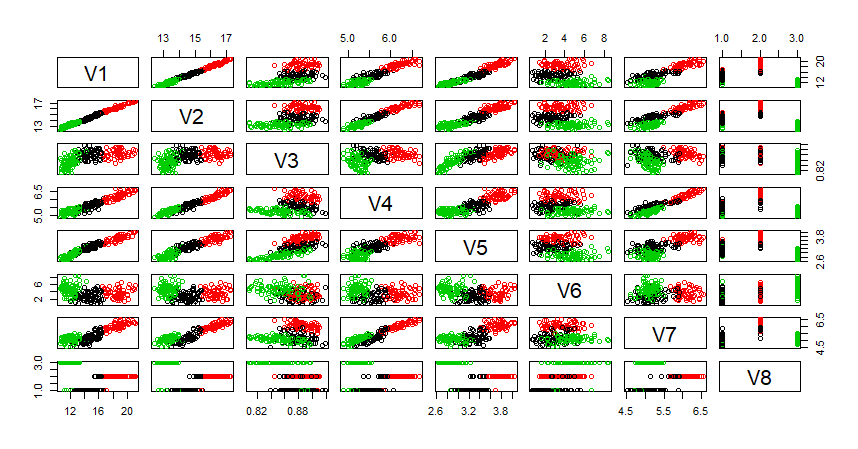
****

Рис. 6. Кластеризация для примера «Seeds»