Министерство образования и науки Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра «Телематика»

ЛАБОРАТОРНАЯ РАБОТА

ПО ТЕМЕ

«Методы кластеризации»

по направлению 02.04.01.02 «Организация и управление суперкомпьютерными системами»

Выполнил:
Студент гр. 13643.1 Титов А.И.
Проверил: Уткин Л.В.

Санкт-Петербург
2019

Оглавление

Постановка задачи

Требуется решить 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 = $\frac{4\pi A}{P^2}$;
   4. длина зерна;
   5. ширина зерна;
   6. коэффициент ассиметрии;
   7. длина колоска.

1 Набор данных «Pluton»

Для кластеризации из набора данных были выбраны только первые два признака. Данные были разбиты на 3 кластера с помощью метода k-средних. Было проведено исследование влияния параметра максимального числа итераций на разбиение. Были использованы следующие значения параметров: 1, 1000, 2000. Был построен график разбиения (Рис. 1). На рисунке представлены результаты трех разбиений по значению изменяемого параметра, соответственно. Как можно пронаблюдать - многие точки были причислены к разным кластерам в зависимости от разбиения.

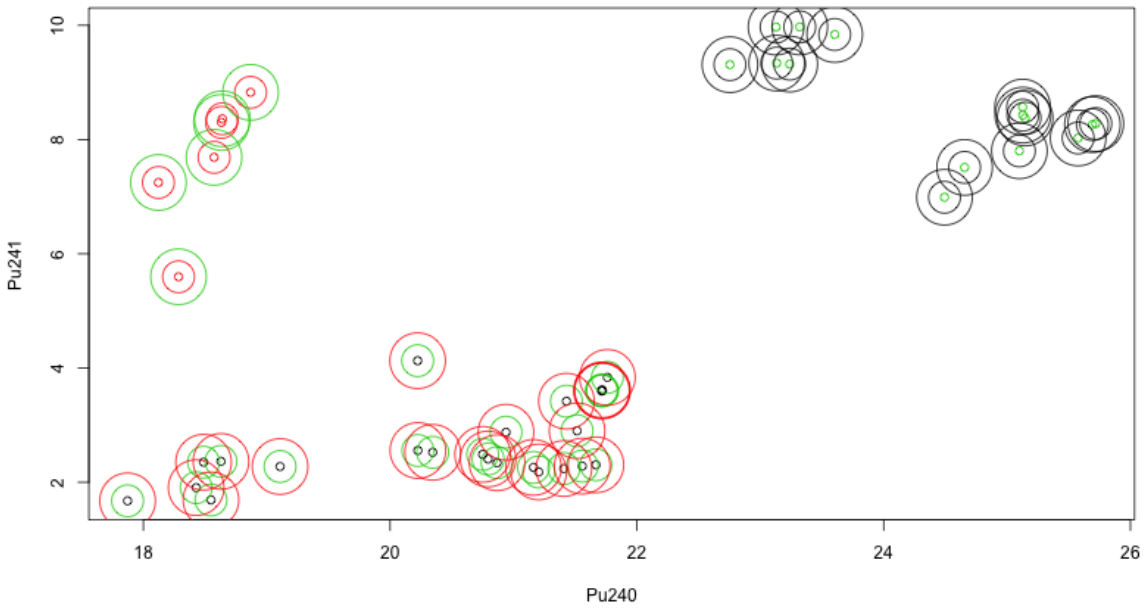

Рис. 1. Кластеризация набора данных «pluton»

2 Сгенерированный набор данных

Был сгенерирован набор данных, состоящий из 3ех кластеров. Каждый кластер вытянут по одной из осей. Была проведена кластеризация методом clara с использованием 2 метрик (manhattan и euclidean) и с использованием стандартизации и без нее. Кластеризация была визуализирована (Рис. 2).

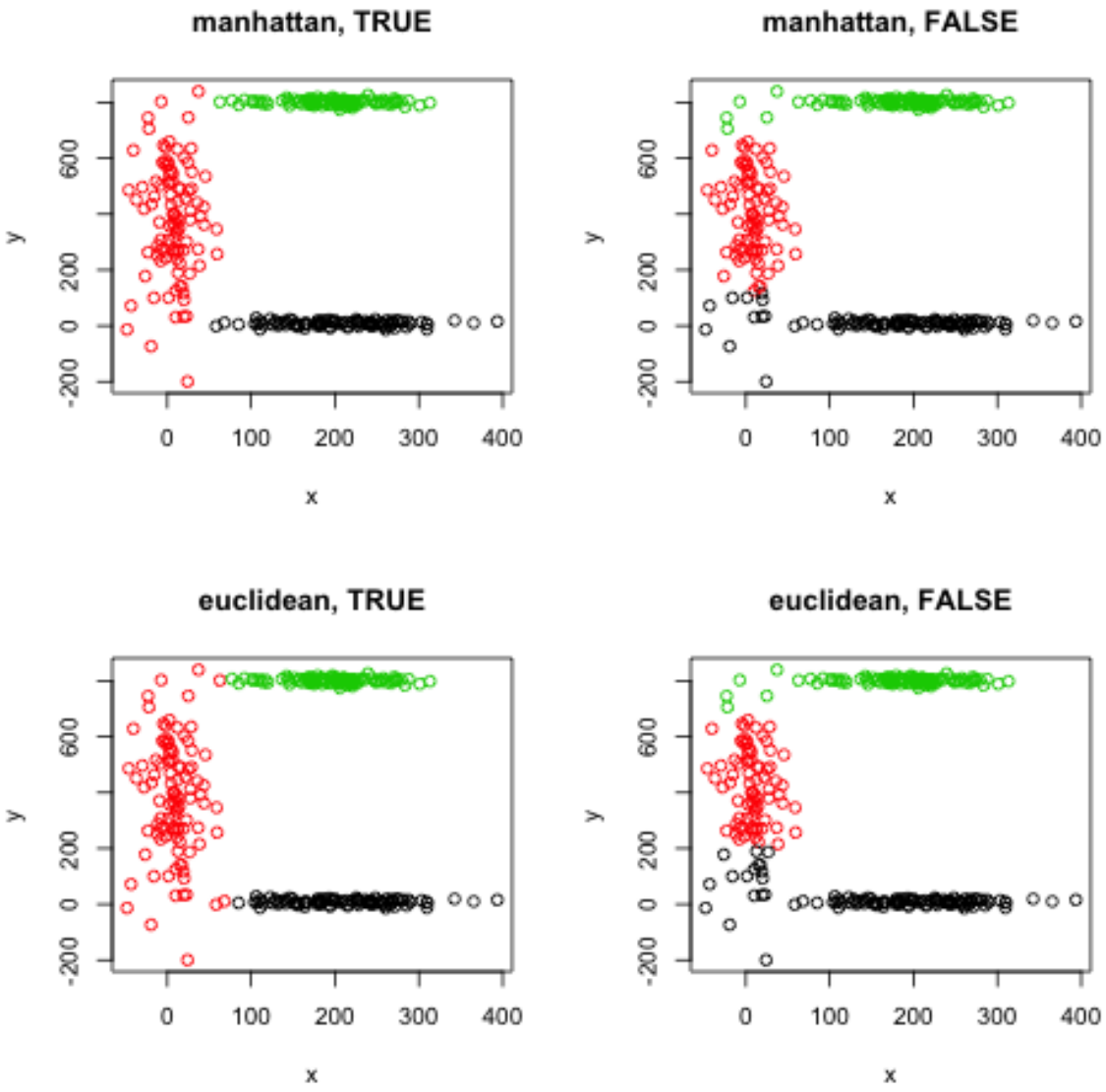

Рис. 2. Кластеризация сгенерированного набора данных

3 Набор данных «votes.repub»

Для данного набора данных была построена дендограмма (Рис. 3), отображающая информацию о близости отдельных кластеров к друг другу и показывающая непосредственно последовательность разделения кластеров.

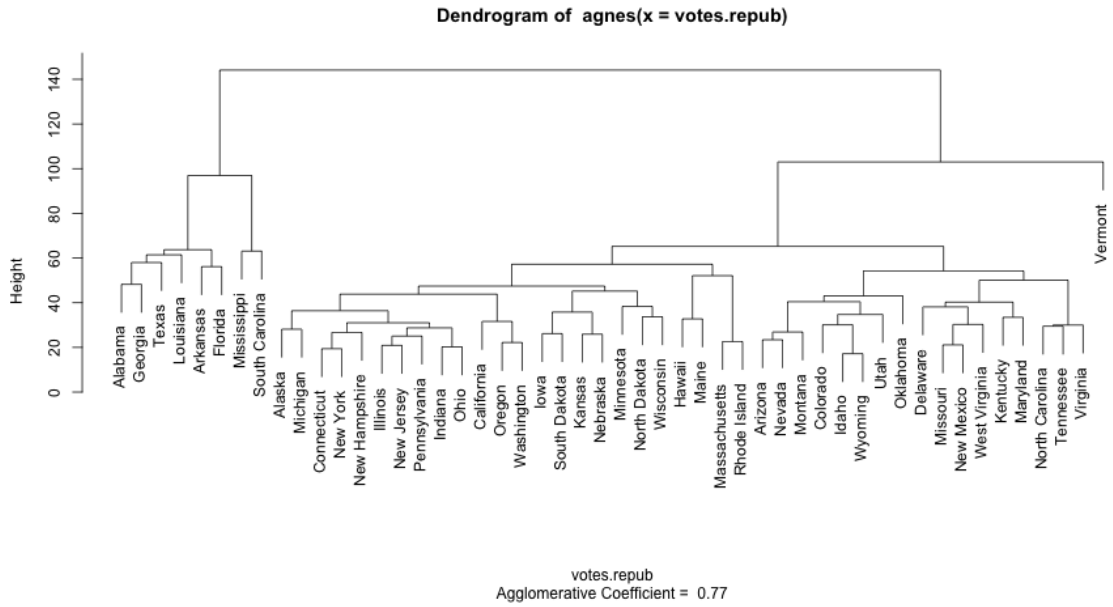

Рис. 3. Дендограмма для набора данных «votes.repub»

4 Набор данных «animals»

Для набора данных была построена дендограмма (Рис. 4). Так как разбиений в данном наборе данных значительно меньше чем в предыдущем, здесь можно пронаблюдать последовательность разделения более детально.

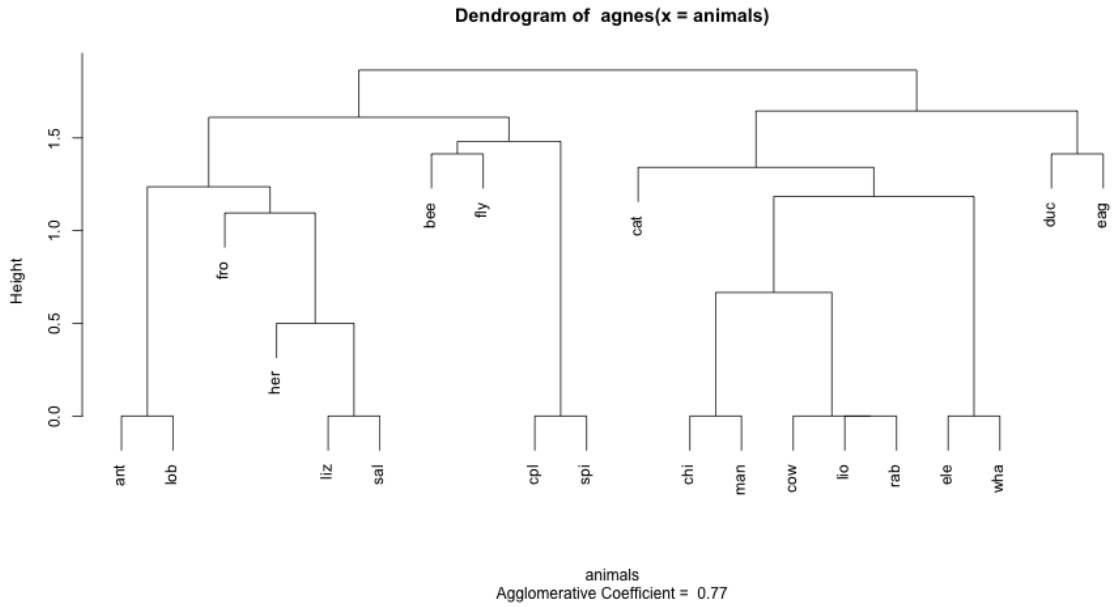

Рис. 4. Дендограмма для набора данных «animals»

5 Набор данных «seeds»

Для набора данных были построены разбиения на три кластера. Разбиения представлены на Рис. 5.

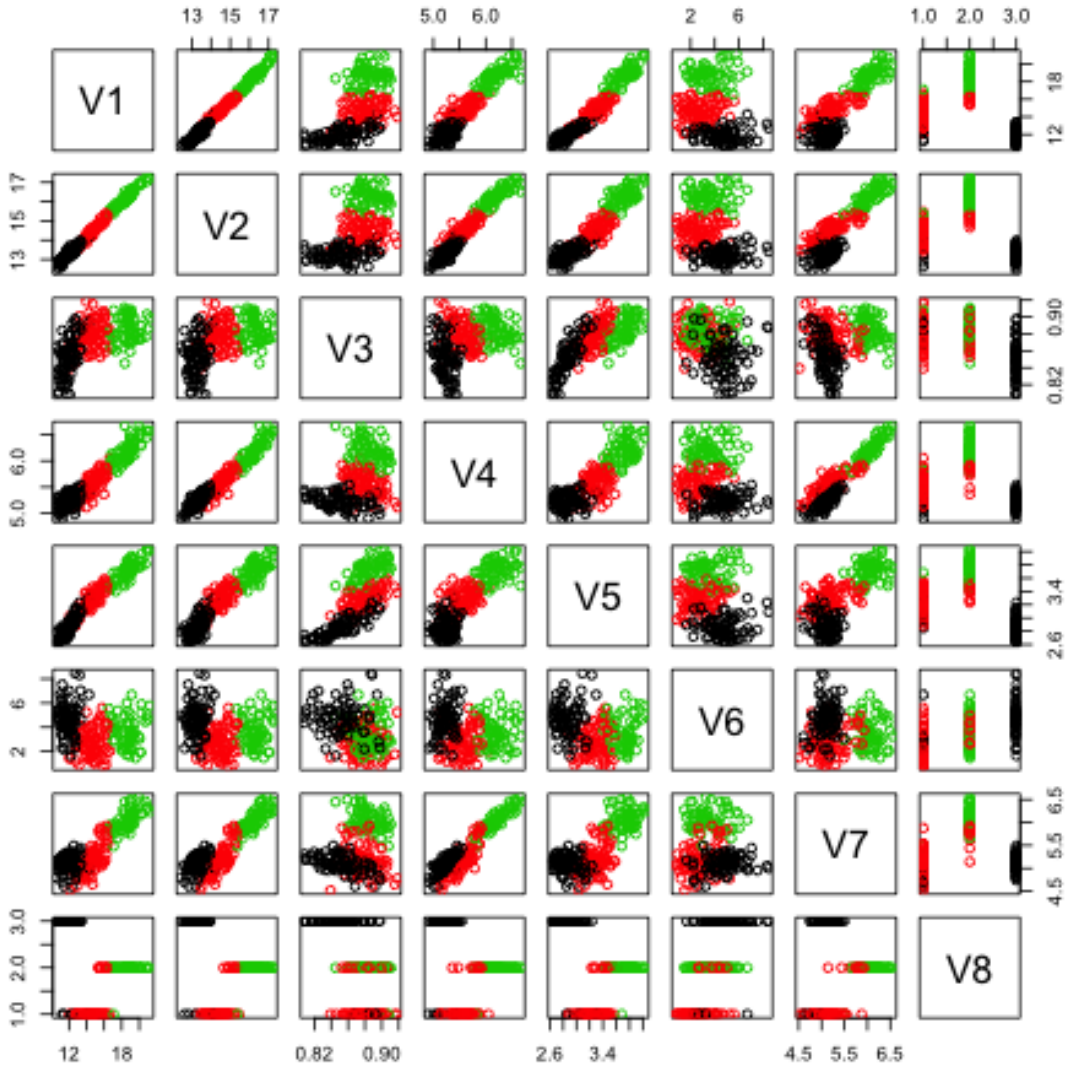

Рис. 5. Кластеризация набора данных «seeds»