Липецкий государственный технический университет

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

Лабораторная работа № 4

# По дисциплине: «Прикладные интеллектуальные системы и

# экспертные системы»

Кластеризация данных

Студент Коретников Н.И.

Группа М-ИАП-23-1

Руководитель Кургасов В.В.

к.т.н. доцент

Цель работы:

Получить практические навыки решения задачи кластеризации фактографических данных в среде Jupiter Notebook. Научиться проводить настраивать параметры методов и оценивать точность полученного разбиения.

Задание кафедры:

1) Загрузить выборки согласно варианту задания.

2) Отобразить данные на графике в пространстве признаков. Поскольку решается задача кластеризации, то подразумевается, что априорная информация о принадлежности каждого объекта истинному классу неизвестна, соответственно, на данном этапе все объекты на графике должны отображаться одним цветом, без привязки к классу.

3) Провести иерархическую кластеризацию выборки, используя разные способы вычисления расстояния между кластерами: расстояние ближайшего соседа (single), дальнего соседа (complete), Уорда (Ward). Построить дендрограммы для каждого способа. Размер графика должен быть подобран таким образом, чтобы дендрограмма хорошо читалась.

4) Исходя из дендрограмм выбрать лучший способ вычисления расстояния между кластерами.

5) Для выбранного способа, исходя из дендрограммы, определить количество кластеров в имеющейся выборке. Отобразить разбиение на кластеры и центроиды на графике в пространстве признаков (объекты одного кластера должны отображаться одним и тем же цветом, центроиды всех кластеров – также одним цветом, отличным от цвета кластеров)

6) Рассчитать среднюю сумму квадратов расстояний до центроида, среднюю сумму средних внутрикластерных расстояний и среднюю сумму межкластерных расстояний для данного разбиения. Сделать вывод о качестве разбиения.

7) Провести кластеризацию выборки методом k-средних. для k [1, 10].

8) Сформировать три графика: зависимость средней суммы квадратов расстояний до центроида, средней суммы средних внутрикластерных расстояний и средней суммы межкластерных расстояний от количества кластеров. Исходя из результатов, выбрать оптимальное количество кластеров.

9) Составить сравнительную таблицу результатов разбиения иерархическим методом и методом k-средних.

Вариант 6 (blobs, random\_state=68, cluster\_std=2, centers=6).

Ход работы:

Добавим в код исходные данные согласно 6 варианту.

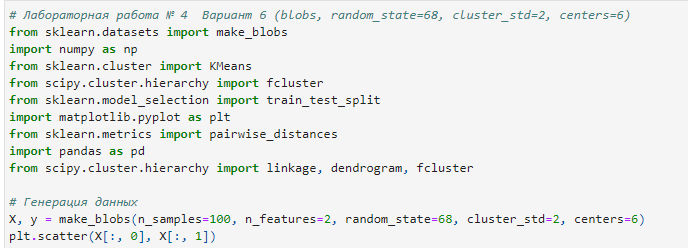


Рисунок 1 – Исходные данные

На основе исходных данных изобразим график.

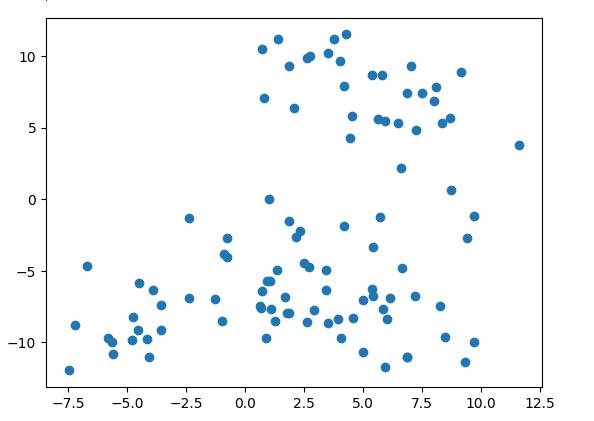


Рисунок 2 – График исходных данных

Через иерархическую кластеризации воспользуемся методами ближайших соседей, дальнего соседа и также методом Уорда. Затем выведем 3 дендограммы для каждого метода.

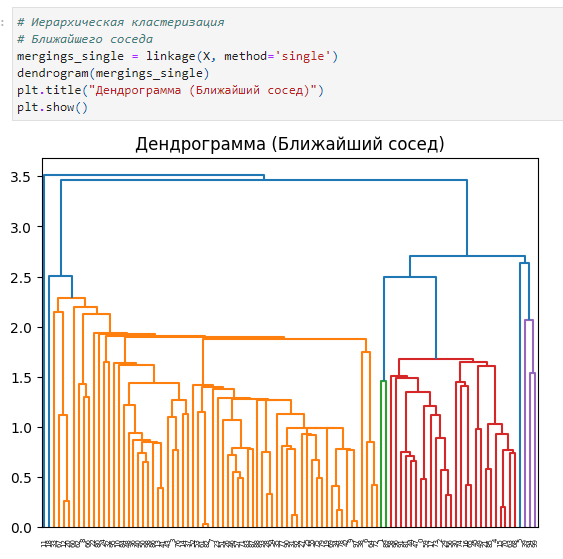


Рисунок 3 – Дендограмма метода ближайшего соседа

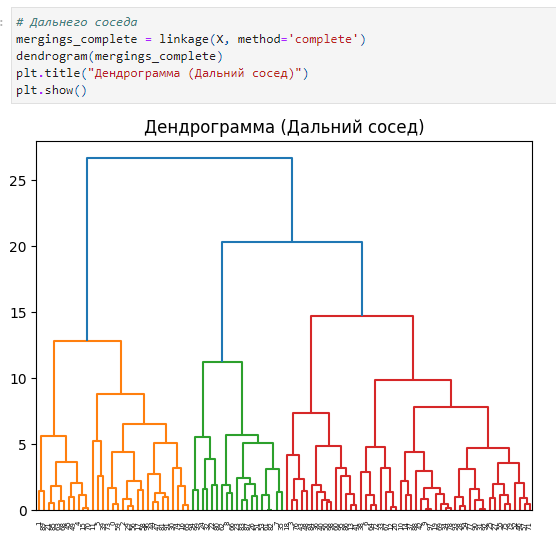


Рисунок 4 – Дендограмма метода дальнего соседа

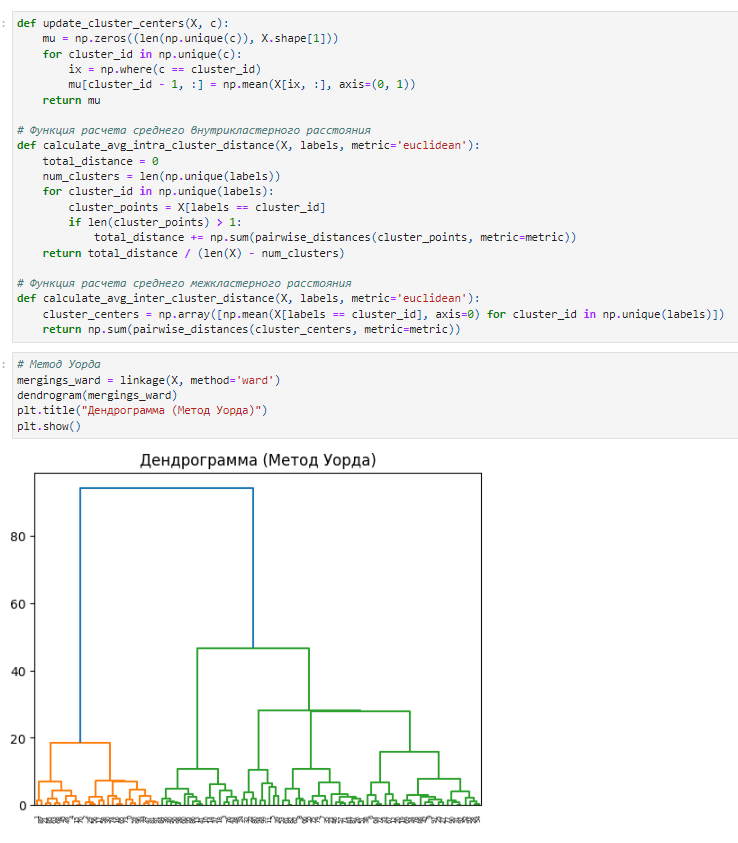


Рисунок 5 – Дендограмма метода Уорда и функция обновления центров кластеров

Определим оптимальную дендограмму из 3 методов.

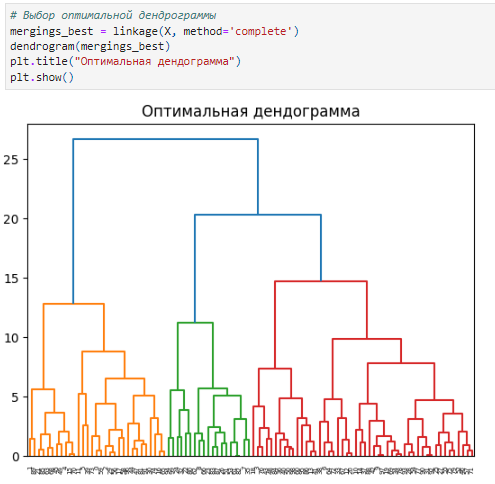


Рисунок 6 – Оптимальная дендограмма

Оптимальная дендограмма это методом дальнего соседа. Определим количество кластеров с иерархической кластеризацией.

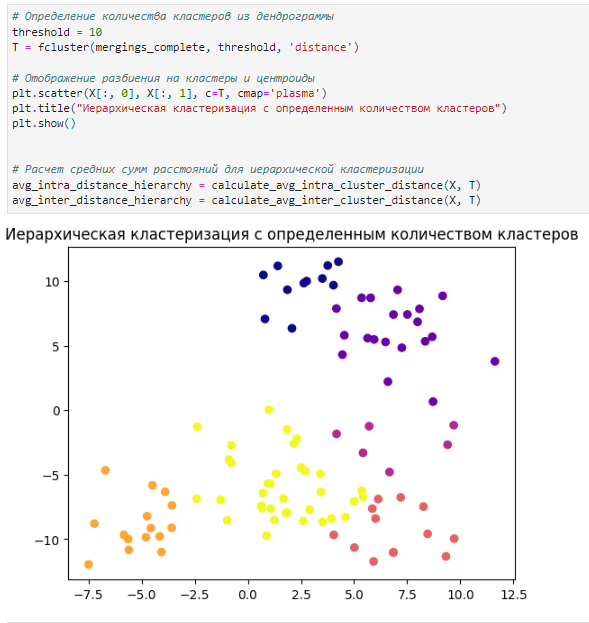


Рисунок 7 – График иерархической кластеризации

Методом локтя определим оптимальное число кластеров.

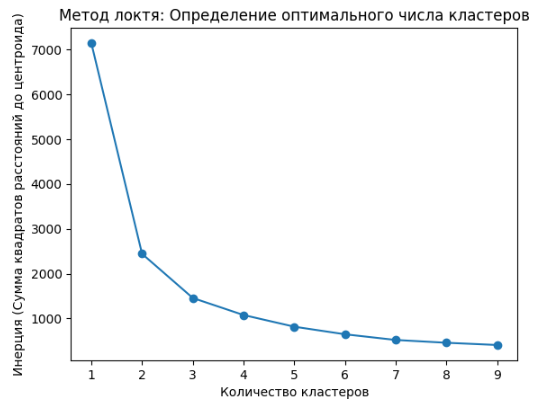


Рисунок 8 – Оптимальное число кластеров методом локтя

Методом k-средних сделаем кластеризацию данных.

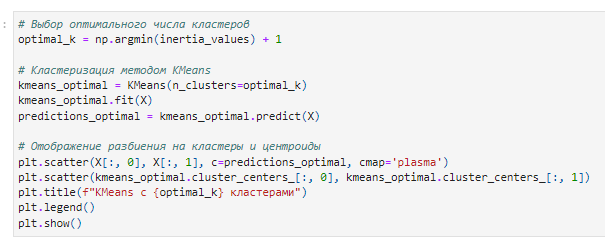


Рисунок 9 – Метод к-средних (k-means)

Через к-среднее изобразим график с кластерами.

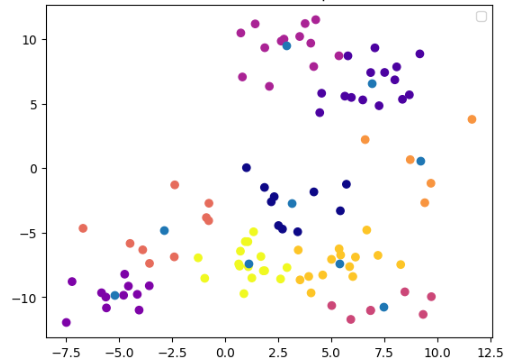


Рисунок 10 – График кластеризации методом к-средних

Сравним расстояние всех методов кластеризации.



Рисунок 11 – Код для создания таблицы и графика для сравнения методов

Выводим после весь результат, показано на рисунке 12.

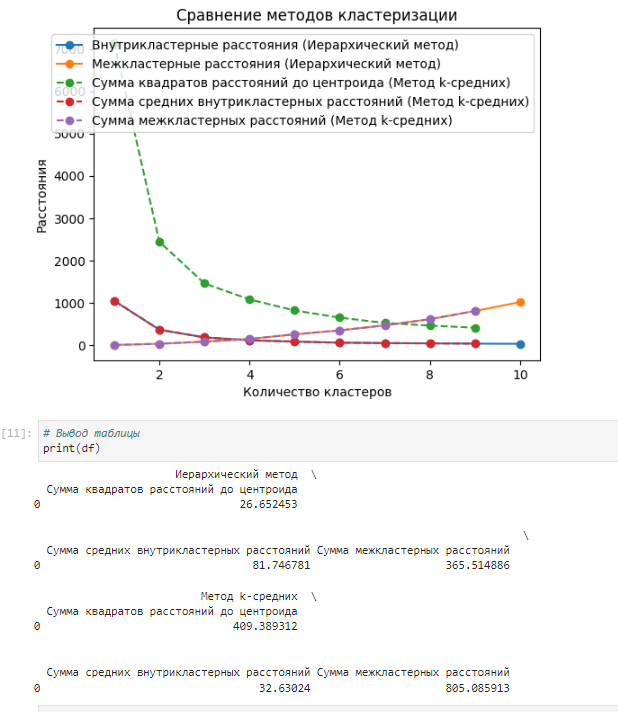


Рисунок 12 – Результат в графическом и в табличном виде

Вывод:

В данной лабораторной работе получили практические навыки решения задачи кластеризации фактографических данных в среде Jupiter Notebook. Научились проводить настраивать параметры методов и оценивать точность полученного разбиения.