*Белорусский государственный университет*

*Факультет радиофизики и компьютерных технологий*

**Лабораторная работа №5**

*«НЕИЕРАРХИЧЕСКИЕ МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА»*

Работу выполнил

Редько Александр

3 курс, 5КБ

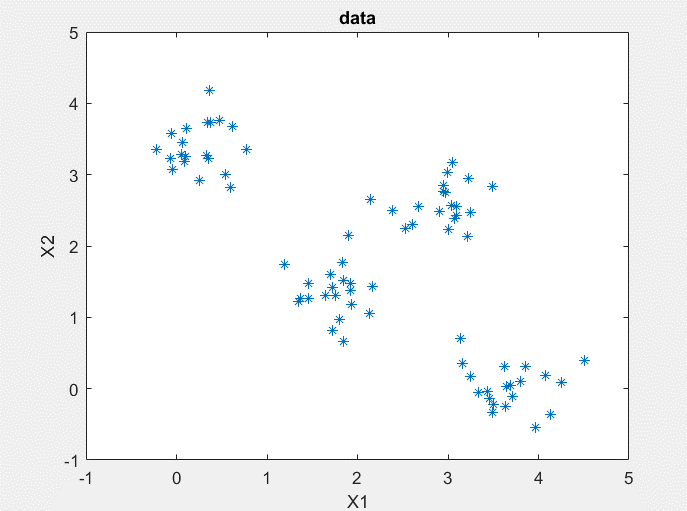
2023

**Цель работы:** Практическое освоение методов неиерархического кластерного анализа многомерных данных на примере метода k-средних.

**Ход работы**

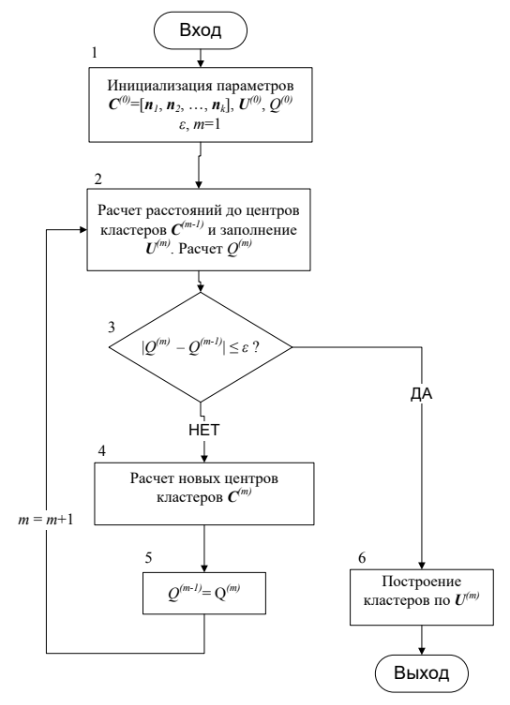
В ходе выполнения лабораторной работы мы, предварительно загрузив исходные данные согласно выделенному варианту, реализовали неиерархические методы кластерного анализа.

После загрузки данных построили их графическое изображение.

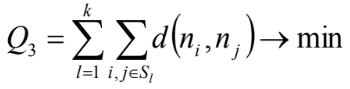


Предварительно предполагаем, что здесь мы имеем дело с четырьмя кластерами.

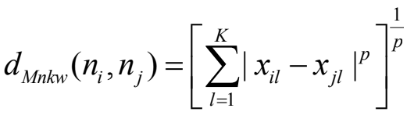
Далее разработали алгоритм кластеризации методом k-средних. Ниже представлена блок схема алгоритма неиерархической кластеризации по методу k-средних:



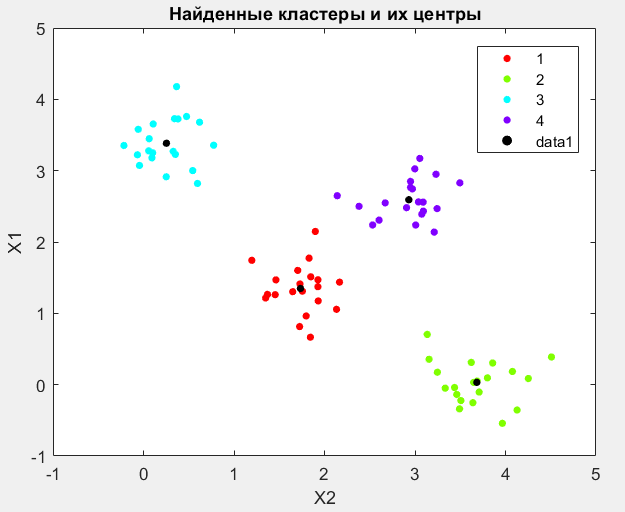
В ходе реализации алгоритма требуется сумму внутрикластерных расстояний между объектами по формуле:



Метрика, которая использовалась для этого согласно моему варианту – Метрика Минковского (p=4).



После выполнения расчетов расстояний и выполнения критерия остановки построили получившиеся кластеры, а также графически изобразили их центры:



**Контрольные вопросы**

1. **В чем заключается задача неиерархического кластерного анализа?**

Задачей кластерного анализа является организация наблюдаемых данных в наглядные структуры – кластеры. Задачей неиерархического кластерного анализа является кластеризация при большом количестве объектов (N>200)

1. **Для каких задач обработки экспериментальных данных используются методы неиерархического кластерного анализа?**

Задачей кластерного анализа является организация наблюдаемых данных в наглядные структуры – кластеры. Методы неиерархического кластерного анализа используются для решения различных задач обработки экспериментальных данных. Вот некоторые примеры:

* + Классификация данных.
  + Сегментация изображений.
  + Анализ генетических данных.
  + Маркетинговый анализ и сегментация клиентов.

1. **В чем суть алгоритма k-средних? Основные шаги.**

Шаг 1. Инициализация начальных параметров метода. Задать: k - количество предполагаемых кластеров, матрицу координат центров кластеров С, матрицу U(0), начальное значение функционала качества кластеризации Q(0) (некоторое большое число) и точность его вычисления (для остановки алгоритма). Установить номер итерации m = 1.

Шаг 2. Рассчитать расстояния от объектов n1, n2, ..., nN до центров кластеров С(m-1), определенных на предыдущей итерации. Заполнить матрицу U(m), исходя из расположения центров С(m-1) вычисленных на предыдущей итерации. Рассчитать значение функционала качества кластеризации Q(m) (с учетом С(m-1)).

Шаг 3. Проверить условие остановки алгоритма. При этом оценивается, привело ли новое объединение объектов в кластеры к существенному улучшению качества кластеризации. Если условие выполняется, то завершить процесс кластеризации. Иначе перейти к шагу 4.

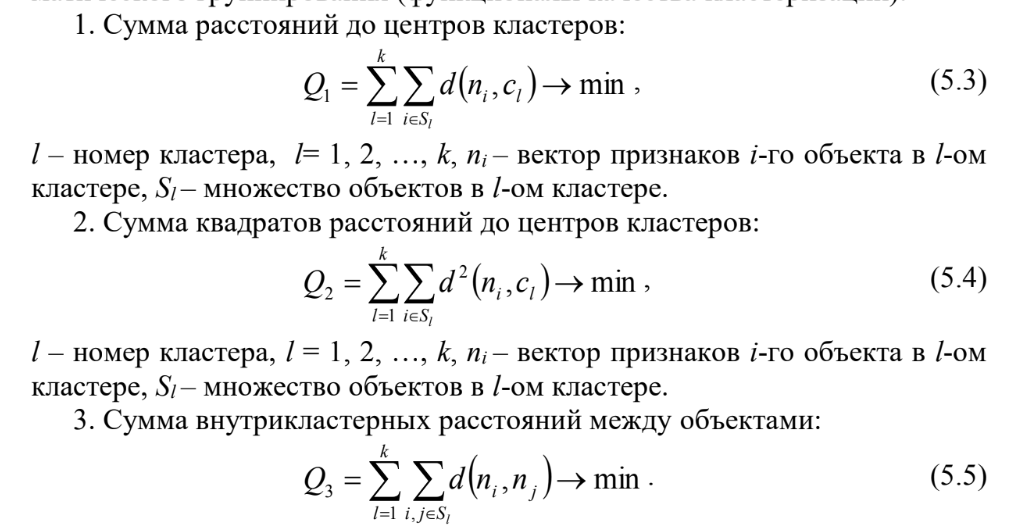
Шаг 4. Рассчитать новое положение центров кластеров С(m) как среднее арифметическое по координатам объектов, входящих в соответствующие кластеры

Шаг 5. Установить Q(m-1) = Q(m) и перейти к шагу 2 (новой итерации) с m = m + 1

1. **Для каких целей в алгоритме k-средних вводится матрица разбиений U?**

Данная матрица вводится для хранения информации о принадлежности объекта к некоторому кластеру. В первом столбце содержатся индексы кластеров, к которым относятся объекты, во втором – расстояния от объектов до соответствующих центров кластеров.

1. **Какие критерии остановки автоматической кластеризации используются на практике?**



1. **Каким образом оценить число кластеров в алгоритме k-средних?**
   * Визуальная интерпретация.
   * Метод локтя (Elbow method).
   * Метод силуэта (Silhouette method).
   * Метод GAP (Gap statistic).
   * Методы информационного критерия (например, AIC, BIC).
2. **scatter, gscatter, min, pdist, std?**
   * scatter: Функция scatter используется для создания точечной диаграммы (scatter plot) в двумерном или трехмерном пространстве. Она принимает входные данные объектов и отображает их на графике в виде точек.
   * gscatter: Функция gscatter (групповая scatter) также используется для создания точечной диаграммы, но позволяет группировать объекты по определенным категориям или меткам. Она принимает входные данные объектов и соответствующие им метки, после чего отображает объекты разных групп в разных цветах или символах на графике.
   * min: Функция min используется для нахождения минимального значения в заданном векторе или матрице. Она возвращает минимальное значение и индекс этого значения.
   * pdist: Функция pdist (pairwise distance) используется для вычисления попарных расстояний между объектами в наборе данных. Она принимает входные данные и метод вычисления расстояния (например, евклидово расстояние, расстояние Чебышева и т. д.) и возвращает матрицу попарных расстояний.
   * std: Функция std используется для вычисления стандартного отклонения значений вектора или матрицы. Она измеряет разброс данных и дает представление о их изменчивости или разнородности.

**Вывод**: в ходе лабораторной работы мы освоили метод неиерархического кластерного анализа k-средних. Реализовали данных алгоритм, а также протестировали его для исходных данных, представленных нам в нашем варианте. Визуализировали полученные результаты.