*Белорусский государственный университет*

*Факультет радиофизики и компьютерных технологий*

**Лабораторная работа №4**

*«ИЕРАРХИЧЕСКИЕ МЕТОДЫ*

*КЛАСТЕРНОГО АНАЛИЗА»*

Работу выполнил

Редько Александр

3 курс, 5КБ

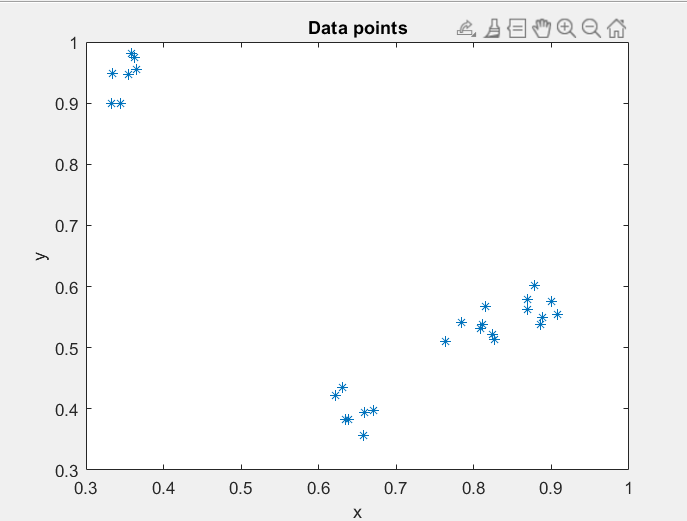
2023

**Цель работы:** Практическое освоение методов иерархического кластерного анализа данных.

**Ход работы**

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

После загрузки данных построили их графическое изображение.



Далее рассчитали расстояния между объектами с помощью методов в выделенном варианте:

* Стандартизированное Евклидово
* Города
* Метрика Чебышева

Выполнили кластерный анализ экспериментальных данных методом иерархической кластеризации по методам связывания согласно выделенному варианту:

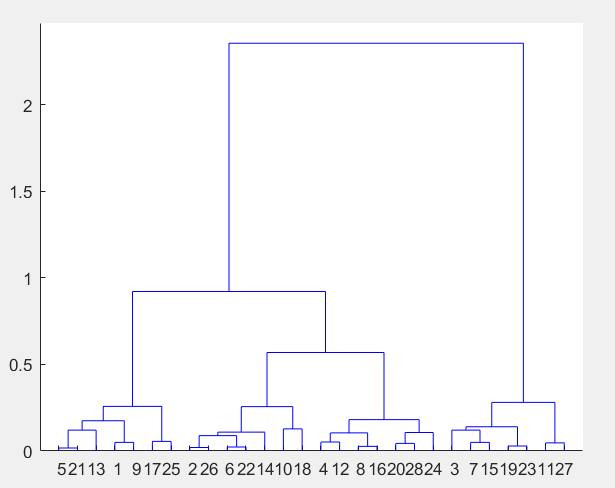
* Метод ближнего соседа
* Метод дальнего соседа
* Центроидный метод

Далее выполнили анализ качества кластеризации с помощью вычисления кофенетического корреляционного коэффициента. Нашли максимальный коэффициент.

|  |  |  |  |
| --- | --- | --- | --- |
|  | стандартизированное Евклидово | города | Чебышева |
| ближнего соседа | 0.9836 | 0.9842 | 0.9753 |
| дальнего соседа | 0.9901 | 0.9900 | 0.9805 |
| центроидный | 0.9902 | 0.9900 | 0.9814 |

Таким образом самым эффективным является: стандартизированное Евклидово, центроидный.

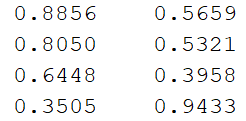
Для наиболее эффективного метода построим дендрограмму:



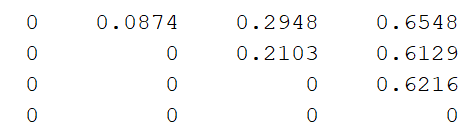
Количество достоверных кластеров: 4.

Рассчитаем центры, внутрикластерную дисперсию, полученных кластеров, геометрическое расстояние от элементов до центров кластеров, расстояния между центрами кластеров.

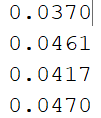
Центры:



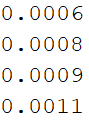
Расстояния между центрами:



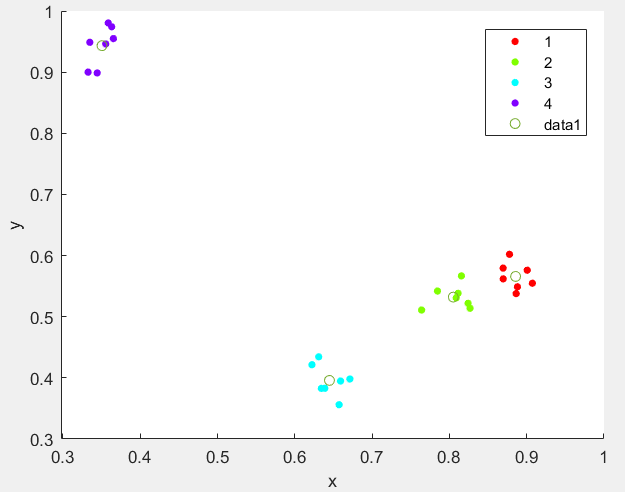
Радиусы:



Дисперсия:



Итоги кластеризации:



**Вывод:** в ходе лабораторной работы мы освоили методы иерархического кластерного анализа. Провели по исходным данным кластерный анализ различными методами, предложенными в нашем варианте. Определили лучший из методов, посчитали дисперсии, расстояния между центрами, а также радиусы кластеров.

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

Кластерный анализ — это метод анализа данных, который направлен на разбиение множества объектов на группы, называемые кластерами, таким образом, чтобы объекты внутри одной группы были максимально похожи друг на друга, а объекты разных групп - максимально различны.

Цель кластерного анализа - найти скрытую структуру в данных и выделить сходство между объектами на основе имеющихся признаков. Этот метод может применяться в различных областях, таких как биоинформатика, маркетинг, социология, финансы и другие, где необходимо классифицировать объекты по определенным признакам.

**2. Для каких задач обработки экспериментальных данных используются методы иерархического кластерного анализа?**

Методы иерархического кластерного анализа широко используются в обработке экспериментальных данных для решения различных задач, например:

* Классификация объектов по сходству: иерархический кластерный анализ позволяет классифицировать объекты на группы по сходству в их свойствах и характеристиках.
* Выявление структур в данных: методы иерархического кластерного анализа могут использоваться для выявления структурных компонент в данных и отделения шума.
* Определение значимых признаков: кластерный анализ может использоваться для определения наиболее значимых признаков объектов, которые могут влиять на результаты эксперимента.
* Анализ социальных сетей: иерархический кластерный анализ может использоваться для анализа социальных сетей и выявления групп людей с общими интересами и поведением.
* Анализ генетических данных: методы кластерного анализа могут использоваться для анализа генетических данных и выявления групп генов, которые связаны с определенными болезнями или характеристиками организма.
* Анализ клиентской базы: иерархический кластерный анализ может использоваться для анализа клиентской базы и выявления групп клиентов с общими потребностями и предпочтениями.

**3. Перечислите основные меры сравнения объектов между собой.**

* Евклидово расстояние.
* Cтандартизированное Евклидово расстояние.
* Манхэттенское расстояние: также известно как расстояние городских кварталов (метрика города). Она измеряет расстояние между двумя точками, используя сумму абсолютных разностей их координат.
* Расстояние Минковского.
* Расстояние Махалонобиса.
* Расстояние Чебышева: это максимальное различие между координатами двух точек, используемое для оценки расстояния между ними.
* Косинусное расстояние: это мера, которая измеряет угол между двумя векторами, заданными координатами точек. Она часто используется для работы с текстовыми данными.
* Коэффициент корреляции: это мера сходства между двумя объектами на основе их корреляции друг с другом. Он измеряет, насколько сильно два объекта связаны друг с другом, и может быть положительным или отрицательным.

**4. Что такое дендрограмма?**

Дендрограмма — это графическое представление результатов иерархического кластерного анализа, которое позволяет визуализировать процесс объединения объектов в кластеры. Она представляет собой дерево, где каждый узел представляет собой кластер, а ветви показывают, какие объекты были объединены для создания этого кластера. Вершины дерева соответствуют кластерам, а расстояния между вершинами отражают степень их схожести. Дендрограмма может быть использована для определения оптимального количества кластеров, которое можно выделить на основе данных.

**5. Что представляют собой иерархические агломеративные методы кластерного анализа?**

Иерархические агломеративные методы кластерного анализа - это методы, которые строят иерархическую структуру кластеров путем последовательного объединения наиболее похожих кластеров, начиная с отдельных объектов и заканчивая одним общим кластером.

В начале каждый объект представляет собой отдельный кластер. Затем, на каждом шаге алгоритма, два наиболее похожих кластера объединяются в один кластер, пока все объекты не будут объединены в один кластер.

В результате получается иерархическая дендрограмма, которая отображает иерархическую структуру кластеров. Каждый уровень дендрограммы представляет собой объединение кластеров на предыдущем уровне, а высота каждой линии на дендрограмме соответствует мере расстояния между объединяемыми кластерами. На основе дендрограммы можно выбрать оптимальное число кластеров для дальнейшего анализа данных.

**6. Что представляют собой иерархические дивизимные методы кластерного анализа?**

Иерархические дивизимные методы кластерного анализа - это методы, которые начинают с одного кластера, содержащего все объекты, и разделяют его на более мелкие подкластеры на каждом шаге. В отличие от агломеративных методов, где шаги состоят в объединении кластеров, дивизимные методы используют процедуру разделения.

**7. Перечислите основные способы связывания объектов в кластеры.**

Основные способы связывания объектов в кластеры (linkage methods) в кластерном анализе:

* Complete linkage (метод дальнего соседа): объекты связываются на основе максимального расстояния между ними.
* Single linkage (метод ближнего соседа): объекты связываются на основе минимального расстояния между ними.
* Average linkage (метод средней связи): объекты связываются на основе среднего расстояния между ними.
* Centroid linkage (центроидный метод): объекты связываются на основе расстояния между центроидами кластеров.
* Ward linkage (метод медианной связи): объекты связываются на основе уменьшения суммы квадратов расстояний между объектами и центроидом кластера при объединении двух ближайших кластеров.

**8. Что такое кофенетический корреляционный коэффициент?**

Мера качества кластеризации, которая используется для оценки соответствия дендрограммы, полученной в результате иерархического кластерного анализа, исходным данным.

Он вычисляется путем расчета корреляции между матрицей расстояний между объектами и матрицей расстояний между соответствующими узлами дендрограммы. Чем ближе значение коэффициента к единице, тем более точная дендрограмма и тем лучше качество кластеризации.

Коэффициент может принимать значения в диапазоне от 0 до 1, где 0 означает, что дендрограмма не имеет никакого отношения к исходным данным, а 1 - идеальное совпадение.

**9. В чем заключаются основные этапы иерархического кластерного анализа?**

* Этап 1. Вычисление матрицы расстояния между объектами D (функция pdist, squareform).
* Этап 2. Связывание или группировка объектов в бинарные иерархические деревья (дендрограммы) (функции linkage, dendrogram).
* Этап 3. Оценка качества кластеризации (функции cophenet).
* Этап 4. Выделение значимых кластеров (функция cluster).
* Этап 5. Визуализация и анализ значимых кластеров (функция gscatter).

**10. Каким образом определить значимое число кластеров?**

Для выделения значимых кластеров можно задать некоторое пороговое значение T меры расстояний сходства (горизонтальная перпендикулярная ось T на дендрограмме). Число значимых кластеров определяется количеством пересечений линии порога T и связей иерархического дерева. Причем каждая из отсекаемых линией порога ветвей дерева будет формировать отдельный кластер. На практике часто выбирают пороговое значение T на основе визуального анализа плотности ветвей построенной дендрограммы.

Альтернативный способ выделения значимых кластеров – метод задания фиксированного числа кластеров. Порогового значение меры сходства T устанавливается в корне иерархического дерева. Затем значение порога T постепенно снижается до тех пор, пока не будет установлено число пересечений линии порога T и связей иерархического дерева равное заданному количеству кластеров.

**11. Для каких целей используются функции MATLAB: pdist, square-form, linkage, dendrogram, cophenet, cluster, gscatter?**

* pdist - функция для вычисления попарных расстояний между объектами.
* squareform - функция для преобразования вектора попарных расстояний в квадратную матрицу расстояний.
* linkage - функция для создания матрицы связей, которая используется для построения дендрограммы.
* dendrogram - функция для построения дендрограммы на основе матрицы связей.
* cophenet - функция для вычисления кофенетического корреляционного коэффициента между матрицей расстояний и матрицей связей.
* cluster - функция для разбиения объектов на кластеры на основе заданного числа кластеров или на основе определенного порога расстояний.
* gscatter - функция для построения графика рассеяния с использованием разных символов или цветов для каждого кластера.