КН-Н921б

Сторожук О.Д.

Лабораторна робота №3

Тема роботи: «Метод просторової кластеризації для даних з шумами DBSCAN»

Мета роботи: дослідження методу просторової кластеризації для даних з шумами (Density-based spatial clustering of applications with noise).

ЗАВДАННЯ ДО ВИКОНАННЯ

Розробити два програмні модулі:

– Програмний модуль генерації псевдовипадкових чисел;

– Програмний модуль просторової кластеризації даних з шумами на основі методу DBSCAN.

Програмний модуль генерації псевдовипадкових чисел повинен генерувати таблицю розміру 3461х2 у діапазоні [1100:1200].

Програмний модуль DBSCAN реалізує метод просторової кластеризації з урахуванням minPts=4 та *ε*=10. Функція знаходження відстані від центру кластеру до об’єкта реалізована за алгоритмом Taxicab geometry.

Також налаштувато minPts та *ε* таким чином, щоб отримати 10% шуму в даних.

ХІД РОБОТИ

Принцип роботи алгоритму DBSCAN полягає у пошуку об’єктів, які належать до єдиного кластеру. Належність визначається гіперсферою із радіусом *ε.* Також умовою існування кластеру є мінімальна кількість об’єктів, що до нього належать – minPts. Відстань між об'єктами визначається за допомогою алгоритмів типу Taxicab geometry (Manhattan distance), Euclidean distance.

Результат роботи для згенерованих даних, які належать до 5 різних кластерів та для параметрів minPts=4 та *ε*=10 наведений на рис. 1.

Chart, scatter chart, bubble chart

Description automatically generated

Рисунок 1 – Результат кластеризації для minPts=4 та *ε*=10

Однак якщо збільшити мінімальний обсяг кластеру до 790, програма розпізнає лише 3 кластери, а інші 1386 об’єктів визначає в якості аномальних (рис. 2).

Chart, scatter chart, bubble chart

Description automatically generated

Рисунок 2 – При збільшенні мінімального розміру кластеру зростає кількість аномалій

При зменшенні minPts до 150 та зменшенні радіуса гіперсфери вдається також спостерігати аномалії. Однак на графіку вони значно ближче до визначених кластерів (рис. 3).

Graphical user interface, chart, application

Description automatically generated

Рисунок 3 – Результат кластеризації для *ε*=1.25 та minPts = 150

ВИСНОВКИ

Під час виконання лабораторної роботи було проведене дослідження методу просторової кластеризації для даних з шумами. Поведінка роботи алгоритму кластеризації була також досліджена для різних його параметрів, таких як площа гіперсфери та мінімального розміру кластеру.