DT584C: Data Mining

Laboratory Task2

Introduction to Clustering Techniques

OBJECTIVES: In this lab, you shall

1. Learn how to apply various clustering techniques
2. Evaluate the performance of your implementation

Note: You may implement your own classification code from scratch or build on available tools.

A dataset containing 200K rows is given. It contains a two attribute dataset ***(x,y)***. Each row is a point in a 2D plane. Your task is to identify the data clusters.

Task1:

You shall apply the k-means algorithm for k=2,3,….
Repeat the above for k-medioid.

Task2:

Repeat the above task using your favourite hierarchical algorithms.
Compare your results (accuracy) with that you found in Task1.

Report:

How many clusters can you see in this data? (What is the plausible number of clusters?). **Motivate your answer using**:

- Mathematical analysis.
- Visual analysis (e.g. plotting XY graph on Excel)

Present a comparison of your results in the two tasks by evaluating the execution results for 1-2 and by examining the implementation of the algorithm for 3-4 below:

1. Accuracy results
2. Performance (execution time)
3. Resource efficiency (Memory consumption)
4. Parallelizability

Show your comparisons in tabular form, present each task's result as a column.