CS565: Data Mining

Programming Assignment 3

Due Date: 12th December, 2007 at 11:59 PM.

Aim of the assignment:
The aim of this assignment is to implement a Density-based Clustering Algorithm and validate its correctness and efficiency using some datasets that we will provide to you. In particular, you need to implement the DBSCAN algorithm described in the section 7.6.1 of the book. You can also download the original paper that describes the algorithm from there: http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf

The datasets will contain 2-dimensional points. The algorithm should take as input the dataset file that contains two dimensional points, the parameter epsilon (ε) and the parameter MinPts. Using these parameters, you should find the core connected points that define the different clusters, the boundary points and the outlier points. Given an epsilon and a certain point *p*, you have to scan the database to find if there are more than MinPts other points with distance to *p* less or equal to epsilon. This step can be performed with a linear scan which means that the running time of your algorithm will be $O(n^2)$.

The output of the algorithm should be the number of clusters, the set of points for each cluster and the set of outliers. You can store this information in an output file that also stores the parameters that used in clustering (ε, MinPts) and the input file. Also, you should plot the produced clusters. Finally, you should produce results with different values of epsilon and MinPts for each dataset.

(Extra Credits: 10%). We suggest finishing with the previous part of the project before you move to this part. You can use an indexing method to speed up the operation to find how many points are in the sphere with radius epsilon around a point *p*. For example, you can use a library with an R-tree implementation. You can find a very good and easy to use R-tree implementation from there: http://research.att.com/~marioh/spatialindex/index.html

Implementation Guidelines:
Implementation platform: Your program should be compile-able and executable in Unix.
Implementation language: It is highly recommended that you use C, C++ and Java.

Input of the algorithm:
1. A database D.
2. The value for the epsilon parameter.
3. The value for the MinPts parameter.

Output of the algorithm:
1. The set of clusters.
2. The set of outliers.

Datasets:
Download the datasets from there: http://www.cs.bu.edu/faculty/gkollios/dm07/P3/
Each of these ASCII files contains a 2-dimensional dataset where each line represents a point.

Deliverables:

1. The source code of your implementation and sufficient instructions on how to compile and run it in Unix.
2. A report that describes briefly the algorithm and the results of some of your experiments. You should provide for each dataset a figure that shows the clusters (points in the same cluster should be distinguished from points in other clusters). To plot these figures, you can use gnuplot or Matlab. Also, you should run a set of experiments with different values of epsilon for the same MinPts and provide a table that shows how many different clusters you detect for different epsilon values. For each dataset, you should find at least three different values of epsilon that give different number of clusters for the same value of MinPts.