Warsaw University of Technology
The Faculty of Electronics and Information Technology

Data Mining (EDAMI)
Project Documentation

Clustering based on density

Author:

Aleksandra Kurdo
Adam Stelmaszczyk

June 13, 2013

Project task

Implementation and experimental evaluation of DBSCAN [**?**] and DENCLUE [**?**] algorithms.

Solution specification

Data set

[**?**]

Set with information about geometrical properties of kernels belonging to three different varieties of wheat was chosen.

Data set attribute information

To construct the data, seven geometric parameters of wheat kernels were measured:

- area A,
- perimeter P,
- compactness C,
- length of kernel,
- width of kernel,
- asymmetry coefficient
- length of kernel groove.

All of these parameters were real-valued continuous.

Algorithms description

DBSCAN algorithm

DENCLUE algorithm

Denclue algorithm is based on the idea that the influence of each data point can be modeled using a mathematical function (influence function). The overall density of the data space can be calculated as the sum of the influence function of all data points. Clusters can be determined mathematically by identifying density-attractors, which are the local maxima of the overall density function.

The Denclue algorithm works in two steps.

Step one:

- It is preclustering step, in which a map of the relevant portion of the data space is constructed. The map is used to speed up the calculation of the density function which requires to efficiently access neighboring portions of the data space.

Step two:

- It is the actual clustering step, in which the algorithm identifies the density-attractors and the corresponding density-attracted points.

Implementation details

As implementation language for the project Java was chosen. Data set is read from a text file, that located in program root folder, and a set of objects that represent points is generated. On this set Dbscan and Denclue algorithms are runs. Other arguments that algorithms need are declared directly in the program files.

Program consists from five packages:

- algorithms package. In this package abstract class ClusteringAlgorithm and inherited from it classes Dbscan and Denclue are located.

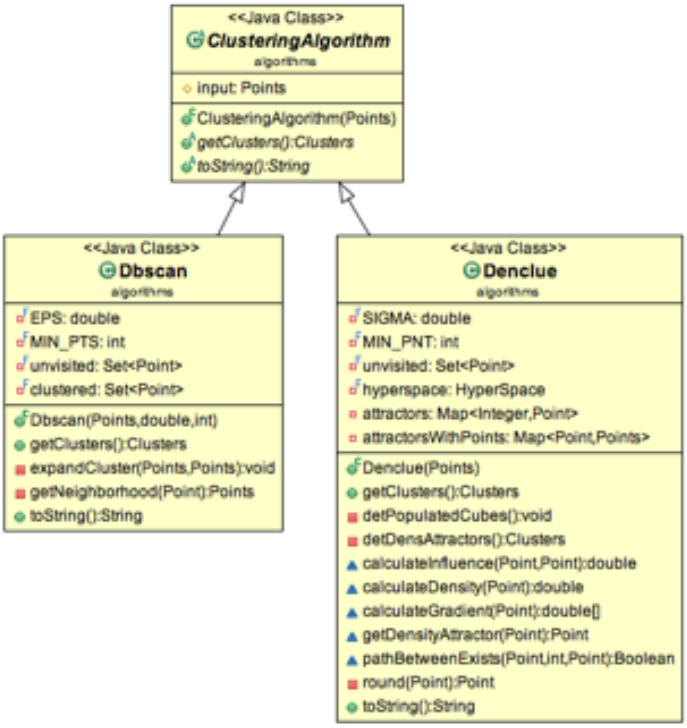

Figure 1: Class diagram for algorithms package

- structures package. Here are located classes Cluster, Point and Points, that used simply to describe set of points and generated clusters.
- scorer package. This package consists of a class Scorer, that used to generate some statistics about created by algorithm clusters.
- visualizer package. Used for the clusters visualization.
- main package. Used for reading the data set from the file and executing algorithms.

0.1 User guide

Figure 2: Class diagram for structures package

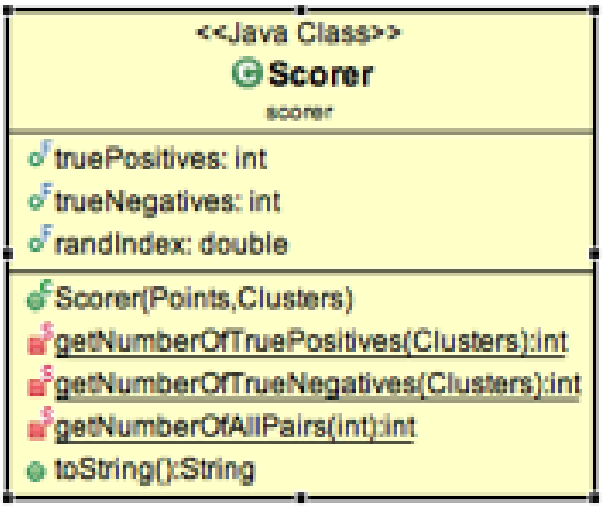

Figure 3: Class diagram for scorer package

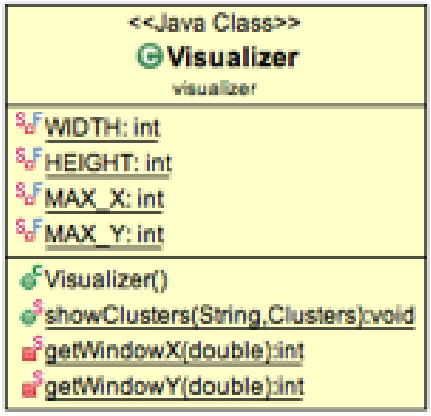

Figure 4: Class diagram for visualizer package

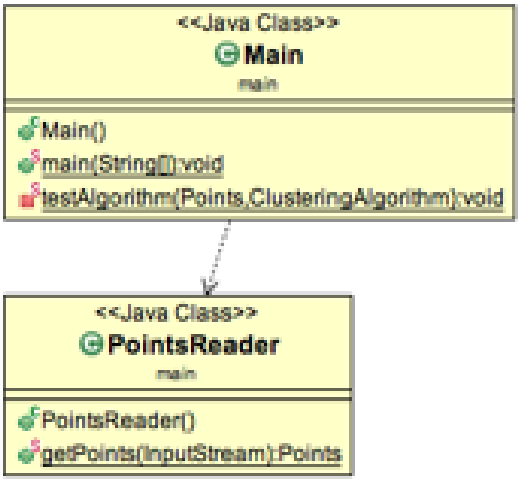

Figure 5: Class diagram for main package