Data Mining (EDAMI)

Project Documentation

“Clustering based on density”

Authors:

Aleksandra Kurdo
Adam Stelmaszczyk

Project task

Implementation and experimental evaluation of DBSCAN [1] and DENCLUE [2] 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:

1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

All of these parameters were real-valued continuous.

Data folder: http://archive.ics.uci.edu/ml/machine-learning-databases/00236/

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.

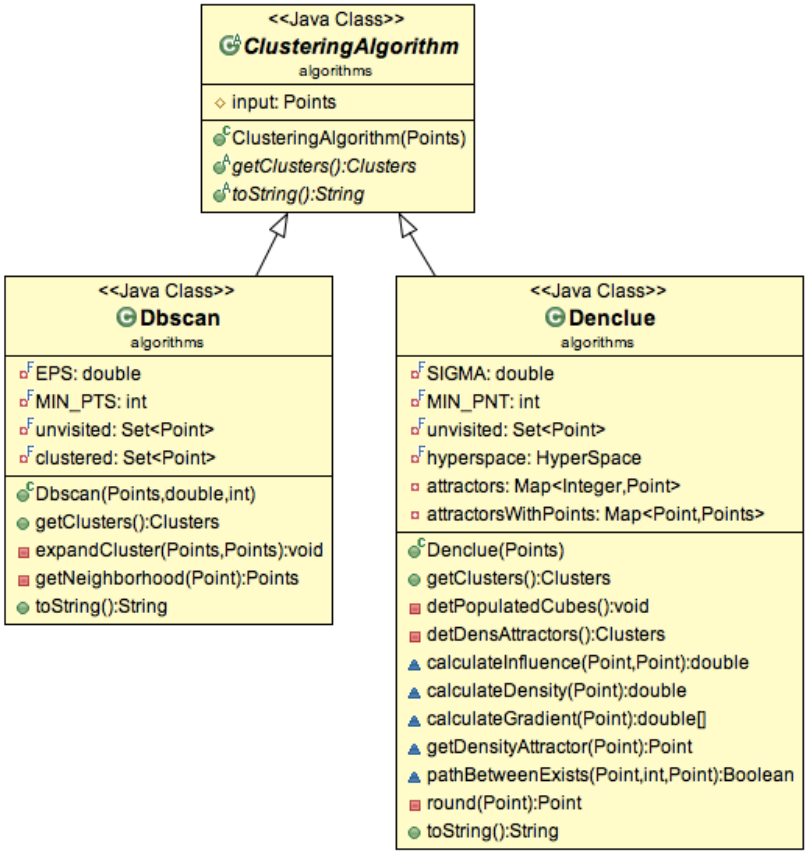
Implementation details

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:

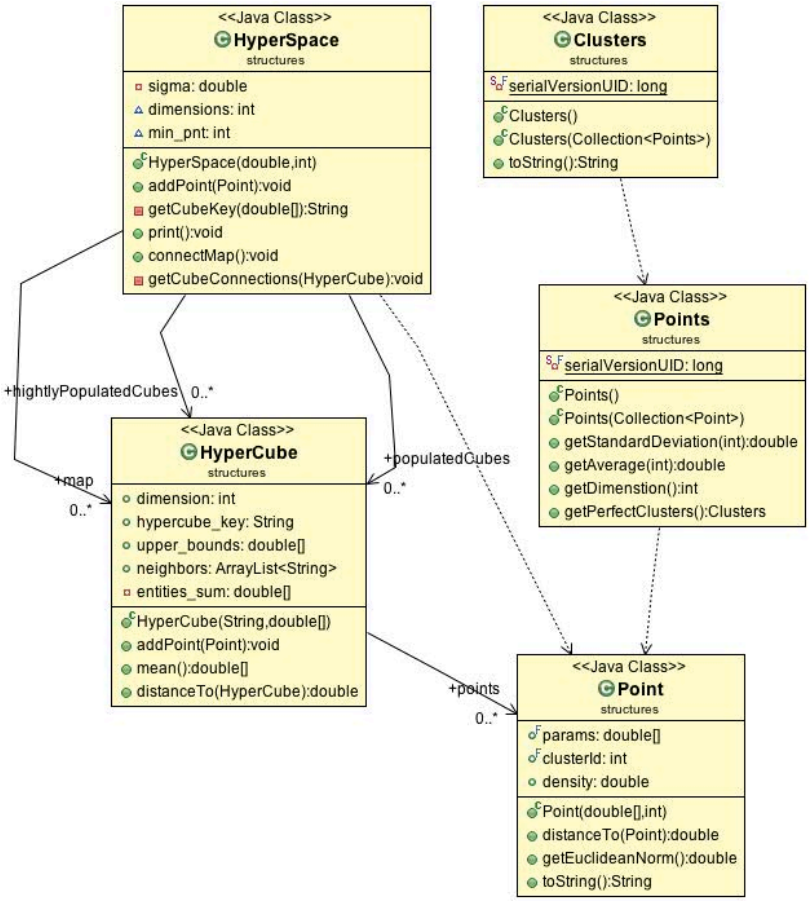
- algorithms package. In this package abstract class *ClusteringAlgorithm* and inherited from it classes *Dbscan* and *Denclue* are located.

  Class diagram:



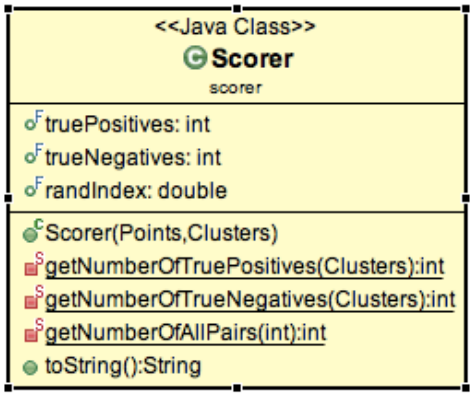
- structures package. Here are located classes *Cluster*, *Point* and *Points*, that used simply to describe set of points and generated clusters.

    Class diagram:



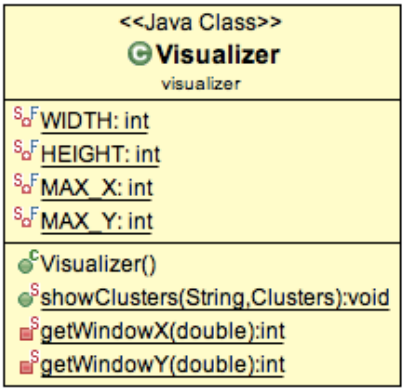
- scorer package. This package consists of a class Scorer, that used to generate some statistics about created by algorithm clusters.

  Class diagram:



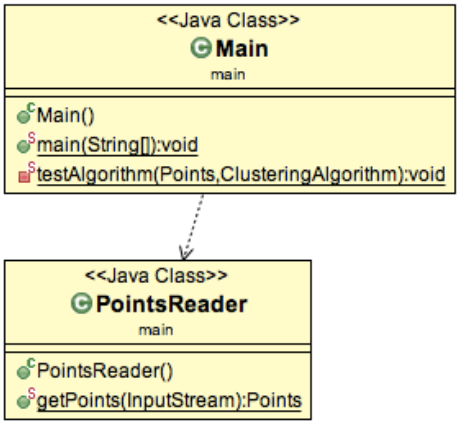
- visualizer package.

  Class diagram:



- main package.

  Class diagram:



User guide

Execute files from edami_clustering.zip.

Create an empty java project in Eclipse IDE and import extracted files to it.

Build the project using Eclipse.

To run the program please add execution rights to the run.sh file by typing in console:

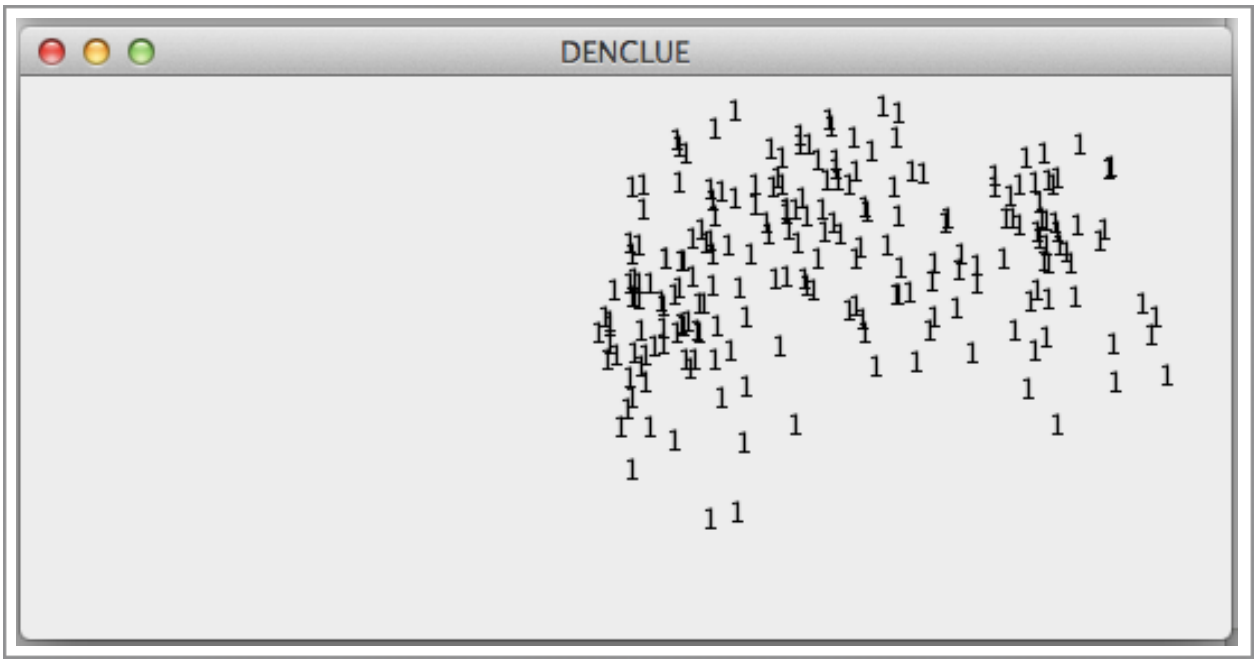
chmod + ./run.sh

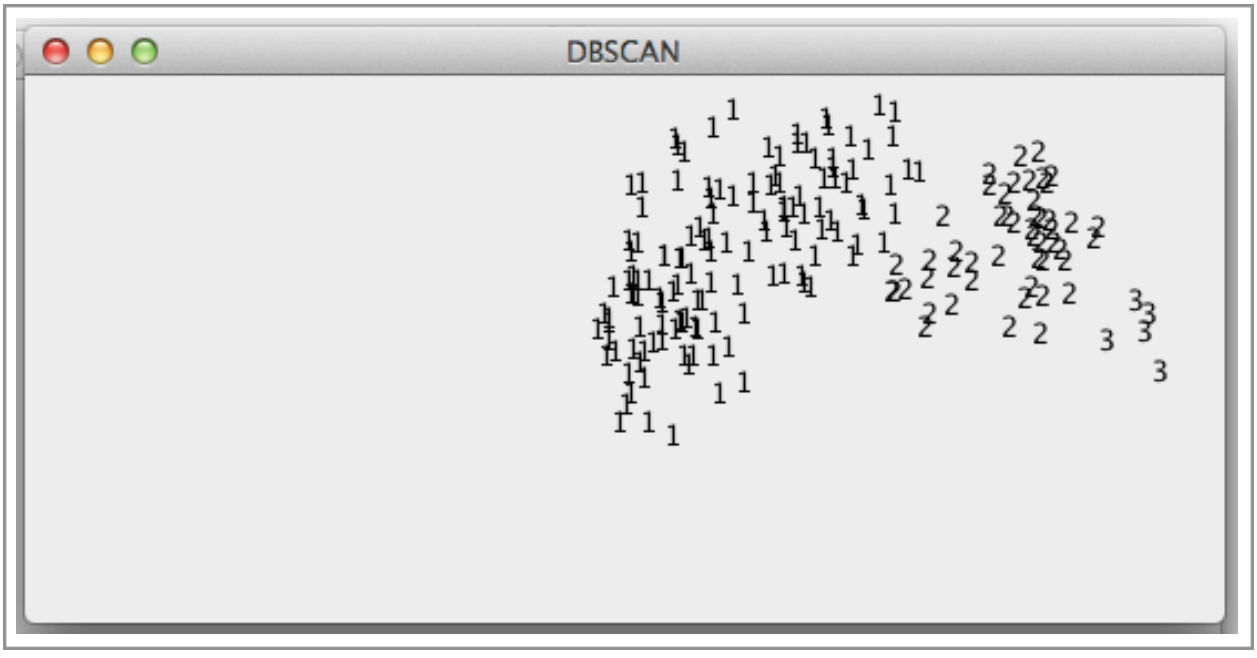
Run program:

./run.sh

Experimentation results and analysis

As a result of execution Dbscan and Denclue algorythms two sets of clusters are generated. This clusters are represented as points in 2D view as a set of points that belong to them. Each point represented as a number of it's cluster. Data set attributes area and asymmetry coefficient are chosen as a dimensions of the view.





Also some statistics for the running algorithms are shown:

```
Standard deviations of parameters:
2.9027633077572266
1.3028455904876302
0.02357308893142152
0.4420073058363384
0.37681405162378667
1.4999729609305972
0.49030891102578644

DBSCAN
3 cluster(s) of size: 134 50 5
True positives: 5402
True negatives: 7150
Rand index: 0.5719753930280246

DENCLUE
1 cluster(s) of size: 210
True positives: 7245
True negatives: 0
Rand index: 0.33014354066985646
```

Where:

$$Rand\ index = (true\ positives + true\ negatives)/(true\ positives + true\ negatives + false\ positives + false\ negatives)$$

Parameters for the algorithms were chosen experimentally. The best results were achieved using for Denclue algorithm:

sigma = 0.7

eps = 3

Dbscan algorithm:

Conclusion

What can be seen achieved by Dbscan and Denclue algorithms results are different. Rand index for Denclue is higher, what means that it's results for this dataset are better.

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". in Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)

[2] Alexander Hinneburg, Daniel A. Keimm, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", Institute of Computer Science, University of Halle, Germany,