Clustering

Project no. 3 [20 p.]

Exercises

Exercise 3.1. [10 p.]

Test the effect of various methods design to detect outliers / anomalies. Consider **Ann-thyroid** data set, that cosists of 7200 instances and 22 variables, is summarized in a table below.

Characteristic	Info	Counts (Percentage)
Numerical variables	V1, V17-V21	6
Binary variables	V2-V16	15
Class Variable	V22	values: 1, 2, 3
Anomaly classes	1, 2	534 (~8%)

Please note that those data are described in the literature (see e.g. [Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. Isolation forest. Data Mining, 2008. ICDM'08.]) to contain known anomaly classes. Hence, we can treat those classes as ground truth and evaluate results of outliers detection algorithms using various measures like AUC, Accuracy, Precision, Recall, etc.

Experiment

1. In [Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. Isolation forest. Data Mining, 2008. ICDM'08.] it was suggested to remove all binary variables in order to compare chosen anomaly detection algorithms.
2. Assume that anomaly labels are unavailable in the training stage.
3. Use
   - DBSCAN
   - HDBSCAN
   - One-class SVM
   - Isolation Forests
   - Local Outlier Factor in order to detect possible outliers / anomalies.
4. Evaluate and compare used methods based on their results using AUC, Accuracy, Precision and Recall.
5. Present your results and conclusions in `Jupyter Notebook` or `knitr` short report.

Exercise 3.2. [10 p.]

The *Clustering Results Repository (v1.1.0)* [https://github.com/gagolews/clustering_results_v1/] provides results obtained using various clustering methods on more than 200 datasets. See also:

- https://clustering-benchmarks.gagolewski.com/weave/results-v1.html#clustering-results-repository-v1-1-0
- https://clustering-benchmarks.gagolewski.com/weave/file-format.html#clustering-results
- https://clustering-benchmarks.gagolewski.com/weave/true-vs-predicted.html

Select one method and prepare report that will include:

- short description of the method;
- comparison of the results of this method to others available in the repository;

- analysis of the strengths and weaknesses of the method, e.g. on selected data sets.

All data sets are available at *Benchmark Suite for Clustering Algorithms - Version 1*[1] [https://github.com/gagolews/clustering_benchmarks_v1].

Note that some data sets are difficult for all algorithms.

[1] M. Gagolewski and others (Eds.), Benchmark Suite for Clustering Algorithms – Version 1, 2020, doi:10.5281/zenodo.3815066

