Project Assignment

Exercise 1 [30 p.]

This project assignment is inspired by investigation conducted in (Bolón-Canedo et. al. 2012). The goal of the experiment is to evaluate various feature selection methods on synthetic data sets with known important and redundant attributes. The experiment that you will conduct includes following steps.

(A) Generate synthetic data.

Propose method that will allow you to generate a synthetic data sets for binary classification purpose. The set should contain small number of relevant features. Then, increase the dimension of the data by adding irrelevant features, redundant features, correlated features, noisy features etc. The important thing here is for you to keep the information about which variables are relevant to the classification task.

You can use or incorporate ready-made implementations of the data generator for classification task e.g. `make_classification` from `sklearn`. But include the short explanation of the data generation process that this function uses in your final presentation.

You may assume that classes in your set are balanced.

(B) Evaluate feature selection methods.

Consider at least three feature selection techniques of your choice but include examples of filter and wrapper methods. Next, apply them to your data in order to detect important features. Finally, assess and compare the quality of each method. Remember, that you know which features are in fact important and take this information into account. You may propose / find a measure that will help you with that. For example, in (Bolón-Canedo et. al. 2012) the following measure was used.

> *A scoring measure was defined in order to fairly compare the effectiveness showed by the different feature selection methods. The measure presented is a index of success* `Suc.` *[...], which attempts to reward the selection of relevant features and to penalize the inclusion of irrelevant ones [...]*
>
> - *The solution is incomplete: there are relevant features lacking.*
> - *The solution is incorrect: there are some irrelevant features.*
>
> $$\text{Suc.} = \left[\frac{R_s}{R_t} - \alpha\frac{I_s}{I_t}\right] \times 100,$$
>
> *where R_s is the number of relevant features selected, R_t is the total number of relevant features, I_s is the number of irrelevant features selected and I_t is the total number of irrelevant features. The term $\alpha = \min\{\frac{1}{2}, \frac{R_T}{I_T}\}$ was introduced to ponder that choosing an irrelevant feature is better than missing a relevant one (i.e., we prefer an incorrect solution rather than an incomplete one). Note that the higher the success, the better the method, and 100 is the maximum.*

(C) Evaluate classification results.

Compare results of SVM and Random Forest classifier on:

- the set including all features (important and irrelevant);
- the set of features selected by each method in step (B);
- the set including only relevant (important) features;
- data projection obtained via selected dimensionality reduction techniques PCA, multidimensional scaling, tSNE.

in terms of accuracy, precision, recall and F1 score.

Repeat steps (A) - (C) few times in order to observe average behavior of considered methods.

Prepare a presentation describing the methodology, results and conclusions.

Literature.

(Bolón-Canedo et. al. 2012) V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, *Knowledge and Information Systems* **34**, Springer, 2012, 483–516.

Exercise 2 [20 p.]

The goal of this experiment is to evaluate the overall quality of selected clustering algorithms on a wide range of benchmark datasets.

Algorithms

In your experiment consider following clustering algorithms:

- K-means
- Genie (with parameter $g \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$)
- Agglomerative Hierarchical Clustering (with single, average, complete and Ward linkage)
- DBSCAN

Datasets

Benchmark Suite for Clustering Algorithms https://clustering-benchmarks.gagolewski.com/weave/suite-v1.html consists of benchmark batteries often used in literature in order to evaluate quality of clustering algorithms. In this project you should use 18 two-dimensional datasets from `sipu` collection, i.e.: **a1, a2, a3, aggregation, birch1, birch2, compound, d31, r15, flame, jain, pathbased, spiral, s1, s2, s3, s4, unbalance.** Datasets and their reference partitions are available through the repository https://github.com/gagolews/clustering-data-v1.

Methdology

Use methodology proposed in [Gagolewski, M., A framework for benchmarking clustering algorithms, SoftwareX 20, 101270, 2022, DOI:10.1016/j.softx.2022.101270, URL:https://clustering-benchmarks.gagolewski.com/].

Detailed description of proposed approach (together with sample code in `Python`) is also given at:

- https://clustering-benchmarks.gagolewski.com/index.html
- https://clustering-benchmarks.gagolewski.com/weave/true-vs-predicted.html
- https://clustering-benchmarks.gagolewski.com/weave/noise-points.html
- https://clustering-benchmarks.gagolewski.com/weave/many-partitions.html

Important

Note that all needed operations are implemented in `clustering-benchmark` package for `Python` (see e.g. https://clustering-benchmarks.gagolewski.com/weave/clustbench-usage.html).[1]

Implementation of algorithms can be found in `sklearn` and `genieclust` packages for `Python`.

Results.

Prepare a presentation describing your results and conclusions.

[1] Only DBSCAN method needs to be consider separately since it requires estimation of ε and MinPTS parameters.

