Knowledge Based Systems Topic Proposal

Marcel Hündorf, Gideon Vogt
mhuendorf@uos.de
gvogt@uos.de

1. PROPOSAL

The topic for the Term Paper is A COMPARISON OF OSDT, GOSDT AND IDS ALGORITHMS. Data sets to be tested on need to be chosen. Therefore we propose different data sets with different amounts of attributes. We also want to have different amounts of cases for each data set. To accomplish that, we plan to randomly choose x cases of a bigger data set, where x will have different predetermined values (e.g. 50, 100, 200, 500 and 1000). For every value of x we plan to conduct experiments with every mentioned algorithm on the resulting smaller data set. The following data sets we want to use are:

- https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King%29
  Chess Endgame Database for White King and Rook against Black King (KRK).
  Number of cases: 28056
  Number of Attributes: 6

- https://archive.ics.uci.edu/ml/datasets/Adult
  Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.
  Number of Cases: 48842
  Number of Attributes: 14

- https://archive.ics.uci.edu/ml/datasets/HIGGS
  This is a classification problem to distinguish between a signal process which produces Higgs bosons and a background process which does not.
  Number of cases: 11000000
  Number of Attributes: 28

- https://archive.ics.uci.edu/ml/datasets/Covertype
  Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data.
  Number of cases: 581012
  Number of Attributes: 54

- https://archive.ics.uci.edu/ml/datasets/Crop+mapping+using+fused+optical-radar+data+set
  This big data set is a fused bi-temporal optical-radar data for cropland classification.
  Number of cases: 325834
  Number of Attributes: 175

The Research Questions we want to clarify with this paper are:

- How well do these algorithms (OSDT, GOSDT, IDS) perform in comparison to each other in regards to computing time, accuracy and understandability?

- How consistent is the classification quality of these algorithms on different data sets?

The implementations with the corresponding papers we plan to use are:

- OSDT: https://github.com/xiyanghu/OSDT
  Paper: https://arxiv.org/pdf/1904.12847.pdf

- GOSDT: https://github.com/Jimmy-Lin/GeneralizedOpt
  Paper: https://arxiv.org/pdf/2006.08690.pdf

- IDS, implemented through pyIds https://github.com/jirifilip/pyIDS
  Paper: https://nb.vse.cz/~klit01/papers/RuleML_Challenge_IDS.pdf

We plan to run the experiments on a local machine, disconnected from the internet or on one of the PCs of the University. A single run is bounded by a time limit. 10 minutes seem to be a reasonable limit.

For assessing the quality of the classifications we will look at the average classification accuracy as well as the f-1 score and the AUC of the classification algorithms on the mentioned data sets. These algorithms are interpretable by design. We don't know of any general metrics to assess interpretability or understandability of them, but we intend to assess the complexity of the models by looking at the number of leaves of the decision trees as well as the number of the rules and the attributes used in the rules for ids, similarly to how it is done in [1]. The computing times will be measured and compared aswell.

2. REFERENCES

[1] ATZMUELLER, M., BAUMEISTER, J., AND PUPPE, F. Semi-automatic learning of simple diagnostic scores utilizing complexity measures. *Artificial intelligence in medicine 37* (06 2006), 19–30.