Interpretability of machine learning models created by clustering algorithms

STU FIIT
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Author: Jakub Janeček
Supervisor: Ing. Jakub Ševcech, PhD.

Motivation

Models separating data into subgroup tend to be hard to interpret. Especially when processing high-dimensional data.

Goals

- Provide interpretation of difference between two data segments based on attribute importance
- Find the trade off between simplicity of explanation and its goodness

Contributions

- Method for identifying subset of attributes that distinguish two data groups
- Metric for evaluating created attribute reductions

$$RM = \left(1 + \left(\frac{n_1 - n_2}{n_1}\right) * w_1\right) * (1 - (r_1 - r_2) * w_2)$$

- n_1 – original size of attr. set
- n_2 – reduced size of attr. set
- r_1 – score of test clf. on orig. data
- r_2 – score of test clf. on red. data
- w_1 ,w_2 – weights for both factors of metric

Used techniques

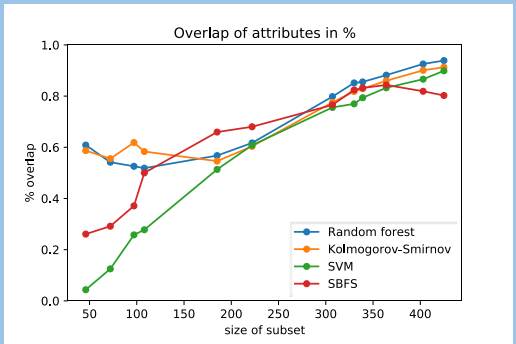
- Logistic regression with L1 regularization for finding the most feasible subset of attributes with our defined metric

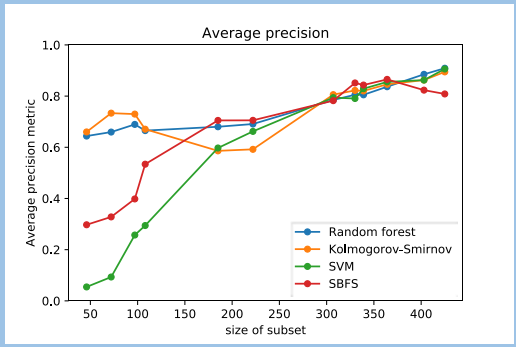
Data and techniques for experiment

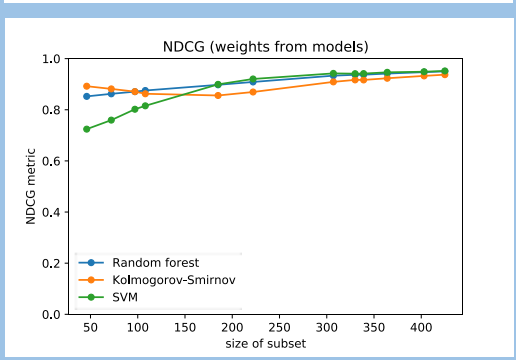
- MNIST dataset as data for clustering
- Topological data analysis (Kepler mapper) as a substitution for clustering algorithm to provide segmentation of data

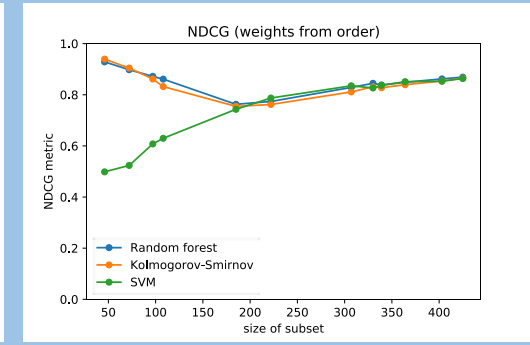
Create reduction → Test reduction → Calculate reduction metric score → Compare reduction results

Did we select important attributes?







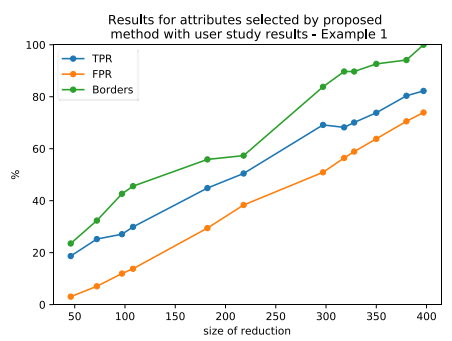


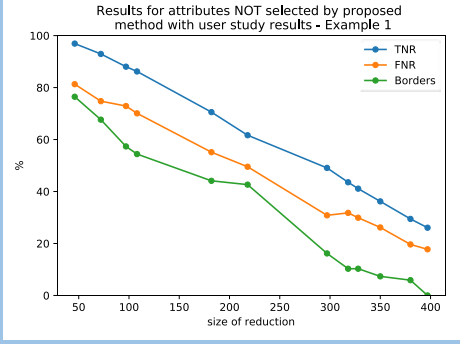
Results

- First part of the results (on the left)
- Four metrics used to measure if the selected attributes with L1 regularization are really important
- Compared against other techniques for feature selection
- Used metrics:
  - overlap of attribute sets (order plays **no** role)
  - average precision (order plays role)
  - two setups of NDCG (order plays role)
- Differences in chosen attributes can be attributed to different underlying approaches for techniques
- Based on these tests (and others mentioned in our work) we concluded we did select important attributes

- Second part of the results (on the right)
- Data from the conducted user study with 21 participants
- Tried to evaluate, if the results our method provides, are useful for easing up the interpretation of clustering model
- Participants were asked to select good and bad attributes for interpreting the differences between two data groups
- Compared the gathered data against the results of our method
- We concluded that we fulfilled this task to some degree, but more can still be done to improve these results.

Did we select good attributes for humans?





User study web application: https://userstudy-dp-xjanecekj.herokuapp.com/