Metalearning

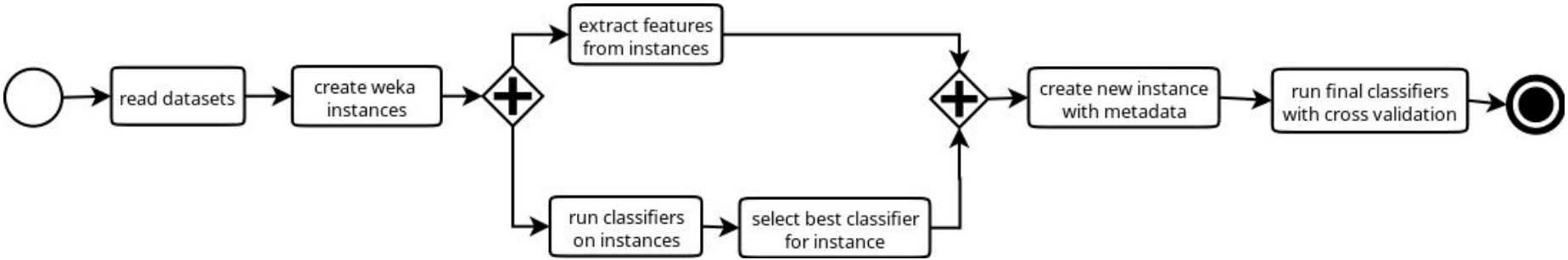
184.702 Machine Learning 2017W - Group 09

Technology

Azure Machine Learning

commons [Math]™

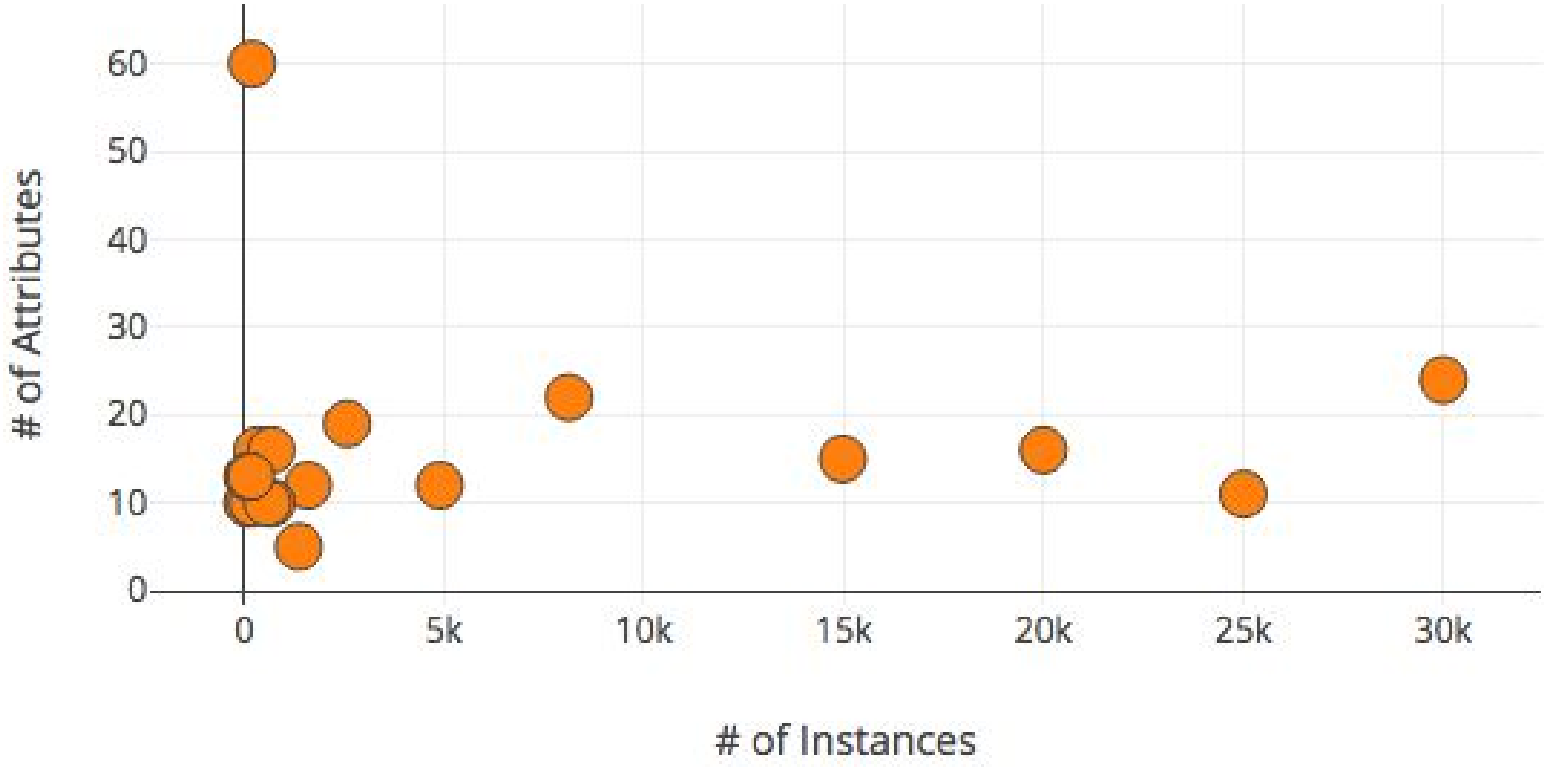
Framework


read datasets
create weka instances
extract features from instances
run classifiers on instances
select best classifier for instance
create new instance with metadata
run final classifiers with cross validation

Datasets

- 20 datasets in total
- Different in
  - Number of attributes
  - Number of instances
  - Number of target classes (binary, multi-class)
  - Type of attributes (numerical, nominal)
  - Type of label column (numerical, nominal)
- Example: Mushroom dataset consists only of nominal attributes (=> no statistical information extraction possible!!)
- Others: Breast Cancer, Credit Screening, Eye detection, Leaf, Letter Recognition, Poker hand, Wine quality, Glass Identification, ...

Datasets


# of Attributes
60
50
40
30
20
10
0
0
5k
10k
15k
20k
25k
30k
# of Instances

Classifiers

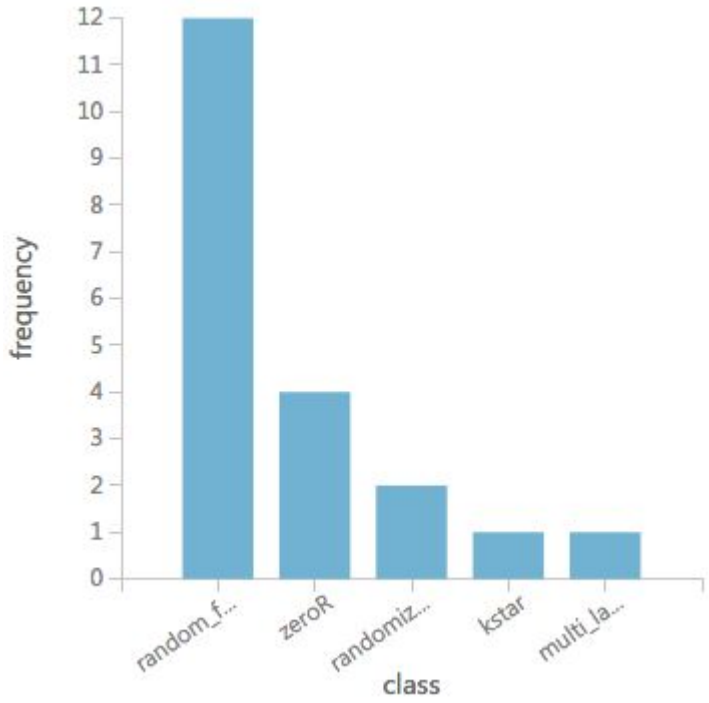
- Bayesian Net
- Multilayer Perceptron
- Random Forest
- KStar
  - Instance based / similarity search / entropy
- Randomizable Filter
  - Uses random classifier / arbitrary filter based on training samples
- RepTree
  - Fast decision tree / pruning based on error reduction / backfitting
- ZeroR Classifier
  - Basic 0R classifier based on mean/mode

Feature Extractors

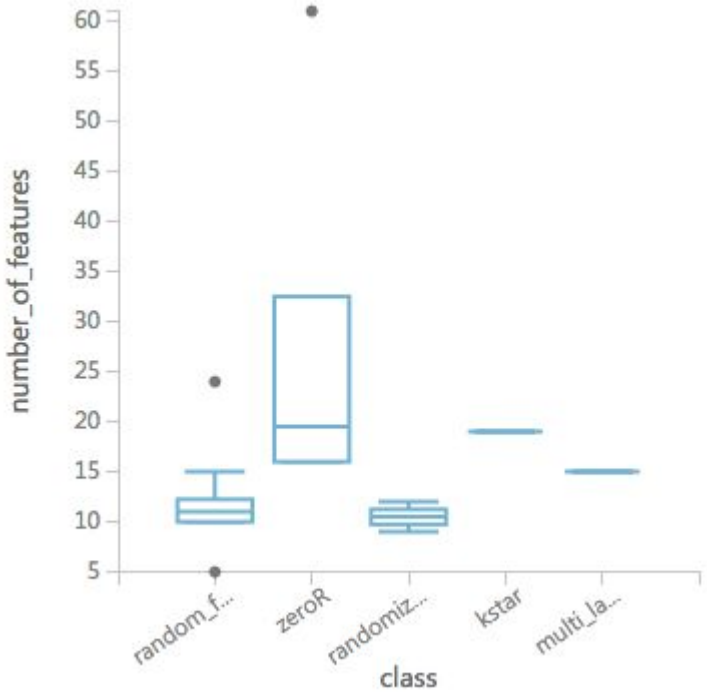
- General
  - Number of classes, features, instances
  - Proportion of missing values
- Statistical
  - Mean of standard deviation
  - Variance, Kurtosis, Skewness, Correlation (both mean and std respectively)
- Information theoretical
  - Entropy
- Model based
  - REPTree size

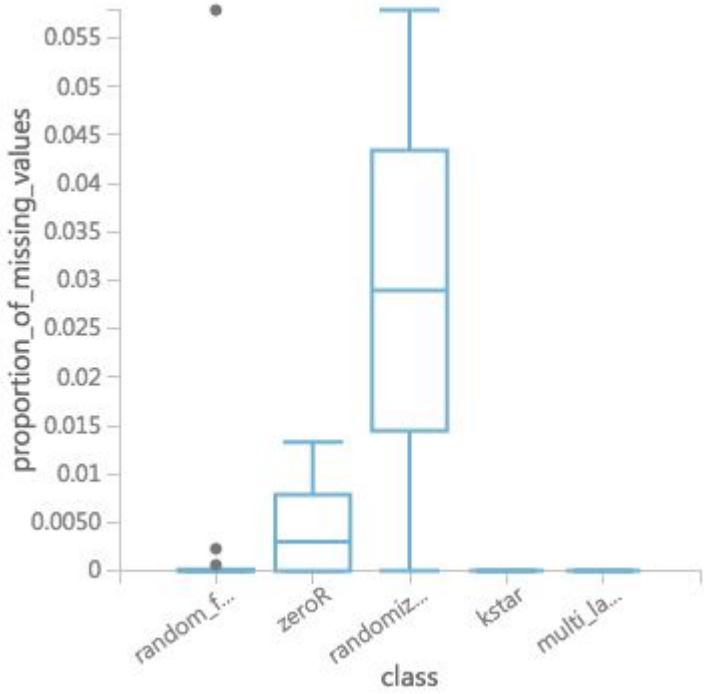
Metalearning dataset

- 15 attributes in total (+ class label)
- Random Forest with highest frequency
- MLP, Kstar only once resp.

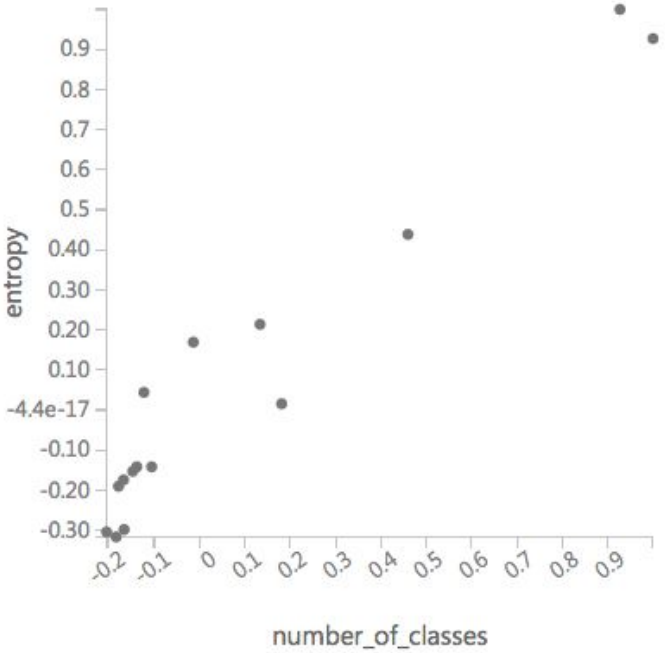


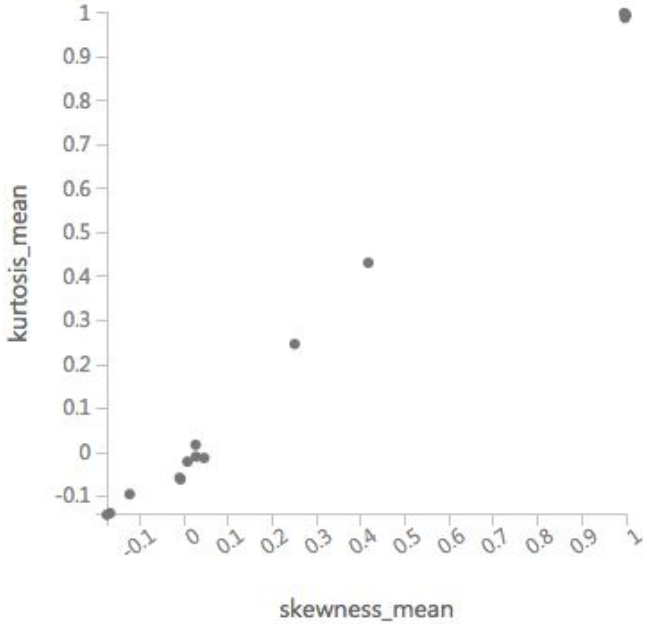
Metalearning dataset





Metalearning dataset

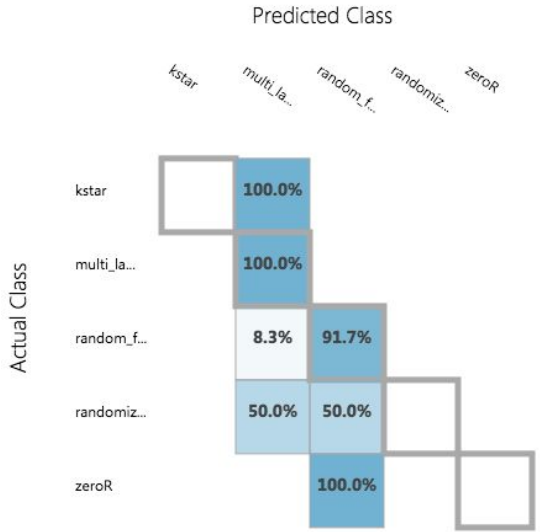




Results

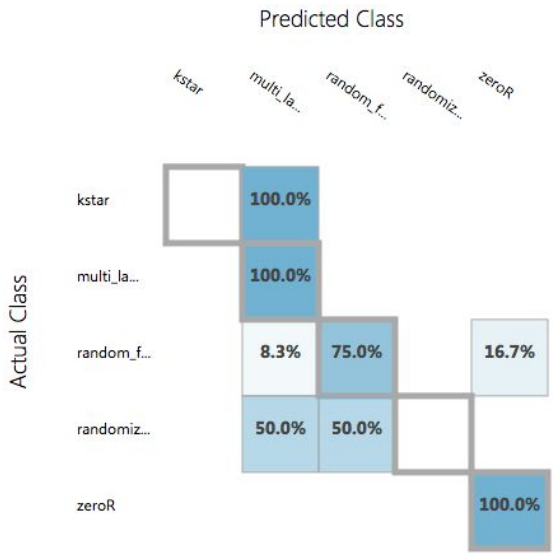
Classifier	Average accuracy	Micro-averaged precision	Micro-averaged recall
Decision Tree	0.84	0.6	0.6
Decision Jungle	0.82	0.55	0.55
Logistic Regression	0.84	0.6	0.6

Results



Overall accuracy	0.6
Average accuracy	0.84
Micro-averaged precision	0.6
Macro-averaged precision	NaN
Micro-averaged recall	0.6
Macro-averaged recall	0.383333

Results



Overall accuracy	0.7
Average accuracy	0.88
Micro-averaged precision	0.7
Macro-averaged precision	NaN
Micro-averaged recall	0.7
Macro-averaged recall	0.55

Summary

- Need even more datasets than used so far
- Runtime even with parallelism around one hour
- Results vary a lot (got worse with more data and then better again)
- WEKA library good choice for Java, but strange behaviour (Exceptions on csv files)

Questions?

Thank you for your attention!