AFAR: automatic feature augmentation ranking

Oliver Neut - o.l.c.neut@student.tudelft.nl
Supervisors: Dr. Rihan Hai, Andra Ionescu

TU Delft

1. Background

Auto ML systems attempt to automate the machine learning pipeline. To increase the model's performance, data augmentation can be used to enrich the existing data.

Efficient and effective **automatic data augmentation** in relational data repositories is a non-trivial task. How to select which tables to join to improve the model's performance?

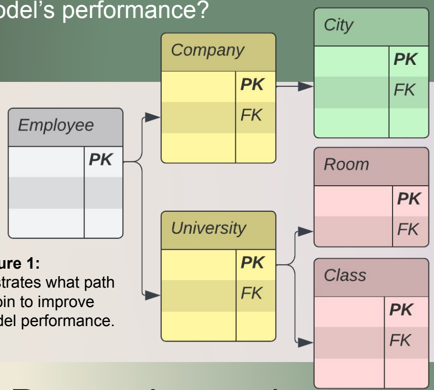

Figure 1: Illustrates what path to join to improve model performance.

2. Research questions:

- What heuristics to select joins make the data augmentation process for XGBoost (decision tree classifier) efficient and effective?
- Define an approach to rank join paths from a relational data repository and validate the:
  - effectiveness (accuracy & depth)
  - efficiency
  - robustness (with other classifiers)

3. Methodology

- Experiment with **feature characteristics**:
  - Categorical data vs numerical
  - Variance, mean, distribution of values
- Look into feature selection **filter methods**:
  - Pearson, Spearman correlation, Information gain, Gini index, Symmetrical uncertainty, ANOVA…[1]

➔ Combine best heuristics to obtain the **AFAR** ranking algorithm

4. Results: AFAR

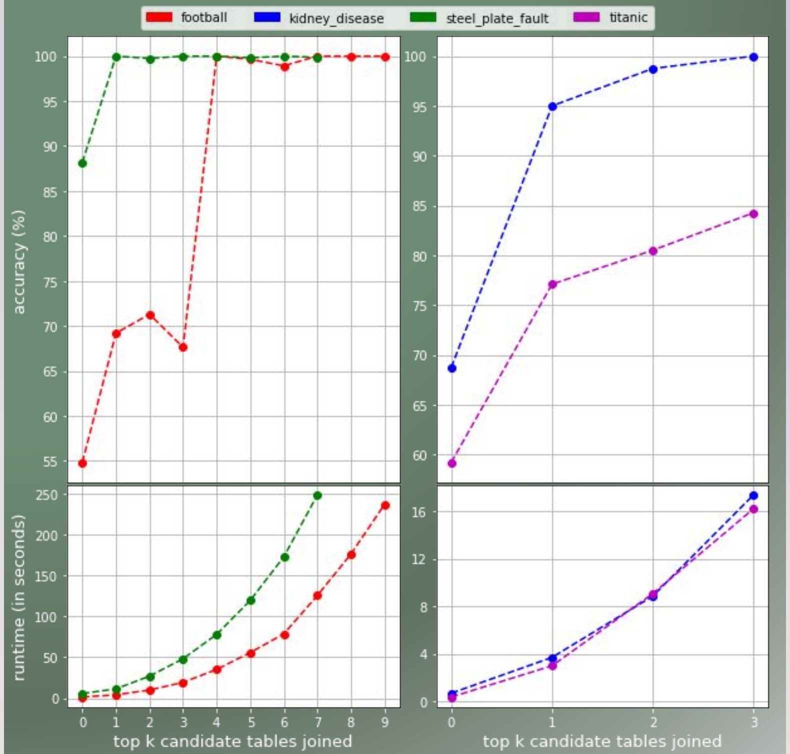

Figure 2: accuracy and runtime of joining the top-k candidate tables determined by **AFAR** Rank_2

4. Results: AFAR

- 2 rank algorithms:
  - Rank_1: Pearson correlation & non-correlation
  - Rank_2: extends Rank_1 with information gain, Gini index and mean unique values score 1/n
- In ¾ datasets → top candidate table ranked 1st

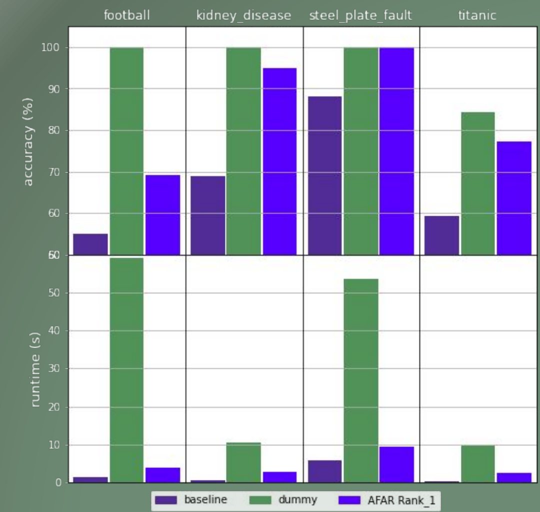

Figure 3: accuracy and runtime of joining the top candidate table determined by **AFAR** Rank_1 vs a baseline and dummy approach

5. Conclusion

- Efficient and effective data augmentation is possible
- To detect good candidate tables select the ones containing features with:
  - high feature-target (base table) correlation
  - low feature-feature (base table) correlation
- The experiments validate that **AFAR** entails:
  - a good accuracy improvement, low max depth
  - low runtime
  - robust against other classifiers

[1] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee, 2015