PCADA: Partial Correlation Aware Data Augmentation

Oskar Lorek

Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft

An Important Problem

Machine learning libraries assume one, flat, tabular data structure as input to the models

Growth in the volume of generated data has led to more unstructured representation of it (Data Lakes)[5]

Increasing impedance mismatch between the data representation and the ML requirements has lead to the rise of importance of data augmentation

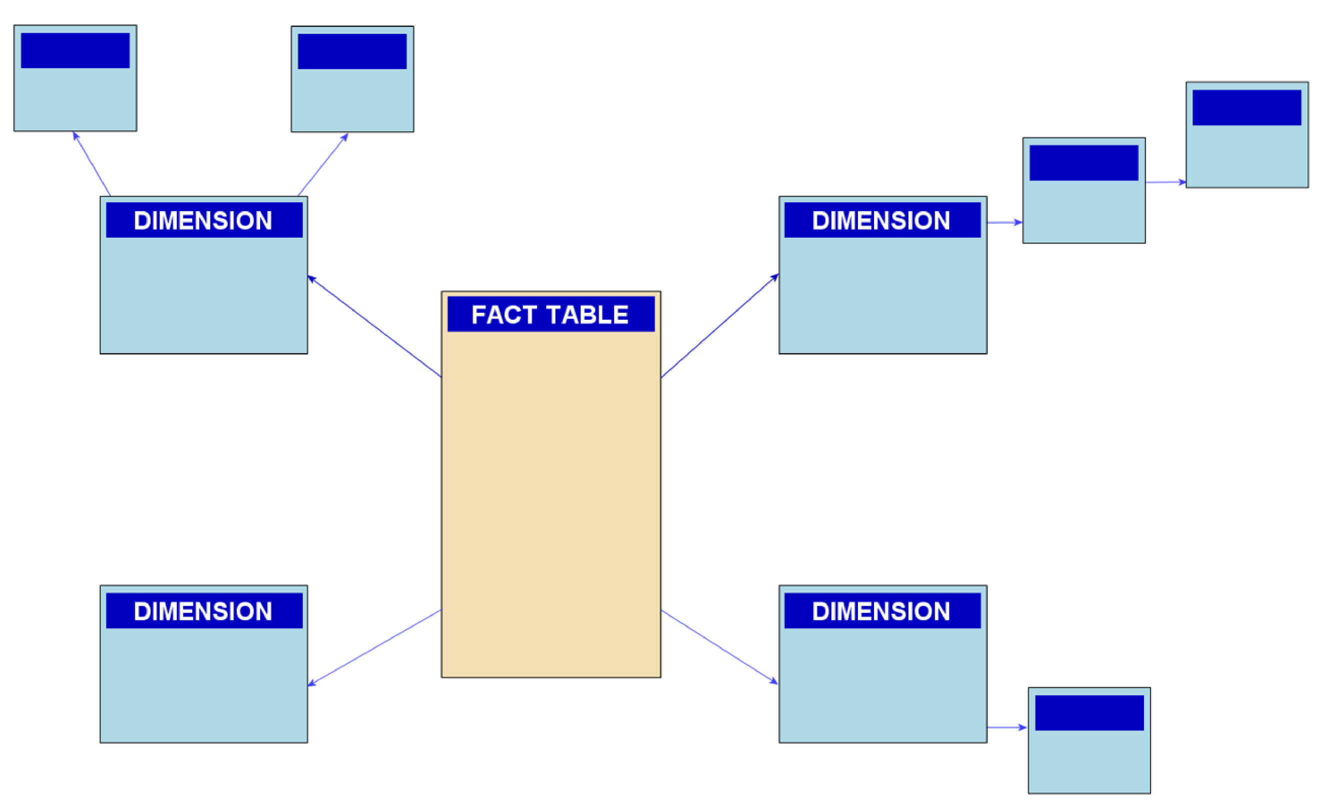

Fig. 1: Typical data schema in modern analytical systems

Choosing too little or undesirable columns leads to low accuracy of the model

An increase in the number of features leads to performance penalty for the ML algorithms

Joins might be expensive and lead to data redundancy, causing even more performance issues[3]

The Knowledge Gap

The problem of selecting features for ML models has been already addressed extensively [4, 1, 2]. However, no major publication examines whether feature selection should account for the type of algorithm that will consume the data. To target this niche, the publication tries to investigate:

What are the characteristics of the optimal features for the random forest classifier?

The Algorithm

Greedy evaluation of importance of neighbouring tables

Sample joins to estimate partial correlation

Partial correlation to decide on whether to join or not

```
function PCADA(target_table, threshold)
    result ← target_table
    frontier ← target_table's neighbours
    while frontier is not empty do
        current ← pop visited
        for all n ∈ current's neighbours do
            s ← result sample join current
            ave_pc ← CALCPC(s, n)
            if ave_pc >= threshold then
                result ← result join current
    return result
```

Fig. 2: PCADA's pseudocode

Why Partial Correlation?

Multi-variable characteristics perform much better than uni-variable characteristics at predicting features' importance

There exists a trade-off between characteristic's effectiveness and the time needed to compute it

Partial correlation performed the best at estimating features' importance

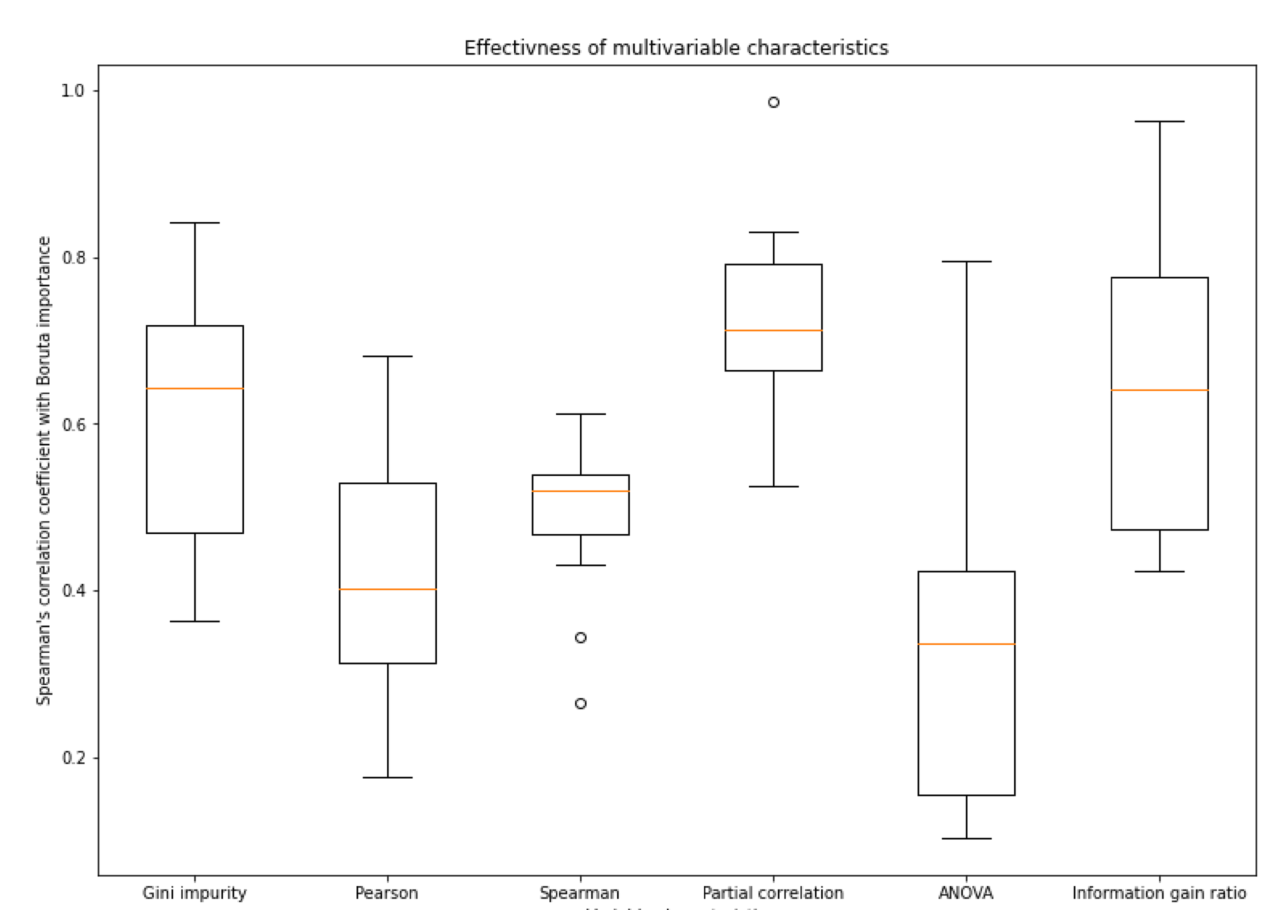

Fig. 3: Effectiveness of multi-variable characteristics

Performance

PCADA achieves similar accuracy to JoinAll approach, while taking significantly less time to train.

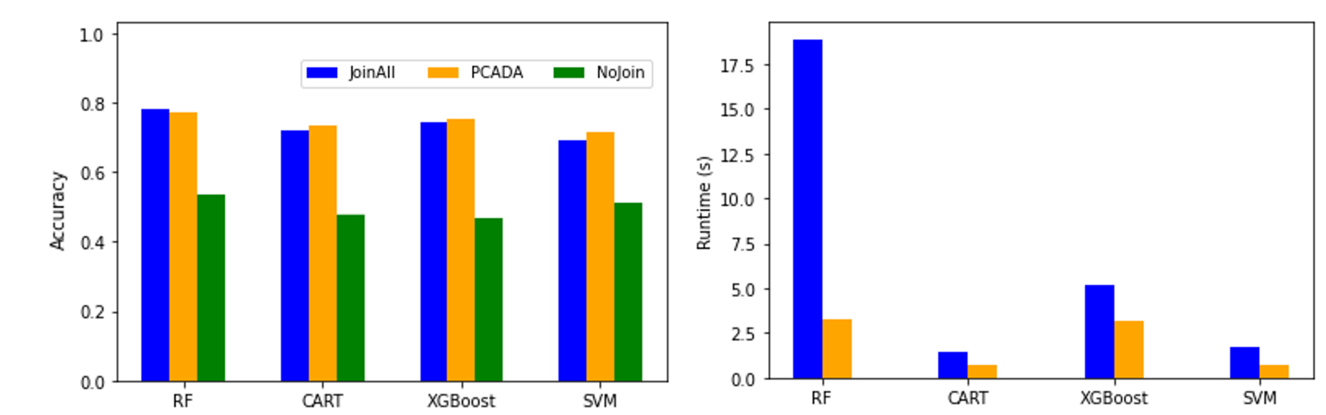

Fig. 4: Accuracy and run time of PCADA for Titanic dataset

Improvements

Non-greedy approach - evaluating on whether to join or not, based not only on the neighbouring table

Determination of optimal sample join ratio - sampling 1% may be infeasible for large datasets

Evaluation of optimal characteristics for other ML models

Acknowledgements

To my supervisor: Andra Ionescu, who introduced me to the problem and provided theoretical support. To the responsible professor: Dr. Rihan Hai, who coordinated the project. To my team-members: Benedict Bien, Eduard Cruset and Oliver Neut, who inspired many ideas implemented in the research.

References

[1] Manoranjan Dash and Huan Liu. "Feature selection for classification". In: *Intelligent data analysis* 1.1-4 (1997), pp. 131–156.

[2] Kenji Kira and Larry A Rendell. "A practical approach to feature selection". In: *Machine learning proceedings 1992*. Elsevier, 1992, pp. 249–256.

[3] Arun Kumar, Jeffrey Naughton, and Jignesh M Patel. "Learning generalized linear models over normalized data". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 2015, pp. 1969–1984.

[4] Jundong Li et al. "Feature selection: A data perspective". In: *ACM computing surveys (CSUR)* 50.6 (2017), pp. 1–45.

[5] Franck Ravat and Yan Zhao. "Data lakes: Trends and perspectives". In: *International Conference on Database and Expert Systems Applications*. Springer. 2019, pp. 304–313.