PCADA:
Partial Correlation Aware Data Augmentation for random forest classifier

Oskar Lorek
Supervisors: Dr. Rihan Hai, Andra Ionescu
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

PCADA: Partial Correlation Aware Data Augmentation for random forest classifier

Oskar Lorek

Supervisors: Dr. Rihan Hai, Andra Ionescu

***Abstract*—Machine learning models require rich, quality data sets to achieve high accuracy. With current exponential growth of data being generated it is becoming increasingly hard to prepare high-quality tables within reasonable time frame. To combat this issue automated data augmentation methods has emerged in recent years. However, existing solution do not focus on specific ML algorithm used for training the data.**

In this paper we propose data augmentation framework designed specifically for the random forest classifier. The algorithm uses sample joins to estimate partial correlation between features in the neighbouring tables and the target column, while controlling for all other features.

Moreover, we show that partial correlation is the most optimal characteristic for determining features' importance for random forest classifier. Apart from it, we demonstrate that PCADA can improve accuracy and run-time in comparison with other baseline data augmentation approaches. Finally, we show that the framework can also be used for other decision trees classifiers (CART, XGBoost) and linear classifier (Support Vector Machine)

***Index Terms*—data augmentation, random forest, decision trees, data lakes, partial correlation, feature selection**

I. Introduction

Nearly all machine learning libraries assumes one, flat tabular data structure as input to the models. However, nowadays this format is different from the way that data is usually stored. Recent growth in the volume of generated data has led to more unstructured representation of it. New terms such as Data Lakes emerges to capture this phenomena[1]. Increasing impedance mismatch between the data representation and ML requirements has lead to the rise of importance of data integration.

Data engineers have to decide which features should be prepared for data scientists. Choosing too little or undesirable columns leads to low accuracy of the model. While an increase in the number of features leads to performance penalty for the ML algorithms. At the same time the actual joins might be expensive and lead to data redundancy, causing even more performance issues[2].

The problem of selecting features for ML models has been already addressed extensively [3], [4], [5]. 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?*

After examining the characteristics of the optimal features for the random forest classifier the paper proposes PCADA, a framework for data augmentation that uses sample join to estimate partial correlation between target feature and columns to be joined, to decide on whether to perform the full join.

Furthermore, the article evaluates suitability of PCADA for other ML models at the same trying to answer the question on whether data augmentation should be ML model aware.

The publication is divided into sections. Related Work presents the current state of knowledge within the area. Proposed data augmentation framework suggests a framework for data augmentation process for the random forest classifier. Evaluation section investigate what are the optimal heuristics for predicting features' importance, then it judges performance of PCADA against other baseline approaches. Furthermore, it shows that PCADA can also be used for other ML algorithms. Responsible research discusses ethical issues related to the research and evaluates reproducibility of the research. The Conclusion section summarizes the outcome of the research in accordance to the research question stated in the Introduction. Further development presents potential disadvantages of the algorithm, as well as the methodology and suggests improvements to tackle mentioned issues.

II. Related Work

A. Feature selection

To examine the usefulness of features, there exists three types of feature selection algorithms: filter methods, wrapper methods and embedded methods.

Filter methods selects the variable irrespective of the model. They try to investigate simple statistical characteristics of the data such as: type of data (e.g. continuous or

categorical), variance, correlation between a feature and the target variable. Filter based methods are particularly fast since they do not require running the ML model, but they suffer in accuracy due to being model and data set agnostic[6].

Wrapper methods try to investigate a subset of variables and evaluate the importance of columns based on running the algorithm and checking the obtained accuracy. Wrapper methods are more computationally expensive than filter methods since they require running the ML model repeatedly. However, the obtained results are much more accurate due to being model and data set related[7].

Embedded methods are a newly proposed class of feature selection techniques which examine characteristics of the data during the ML model training process. In this sense they *embed* the feature selection process within the training. Embedded features try to take advantages of both of the filter method and wrapper methods selection process[8].

Feature selection process for random forest classifier has been dominated by the Boruta algorithm. Its high effectiveness, combined with high efficiency[9] may have led to low number of newly proposed feature selection algorithms in recent years[10]. One can say that Boruta has become a standard practice for feature selection for random forest classifier.

Despite its high efficiency Boruta is not suitable for the data augmentation process. The algorithm is a wrapper method and as a consequence requires access to the data after the join. On the other hand, the data augmentation process has to investigate the importance of columns before the join. Thus, the setting limits us to the usage of characteristics proposed by filter methods.

Irrespective of the above mentioned fact, we can still use Boruta in the research to find the characteristics of features suitable for random forest. The algorithm gives numerical value to the feature importance. Thus, the correlation of the importance of features and its characteristics (for example: variance, correlation with target) can be examined.

B. Data augmentation

The problem of evaluation of the best possible join paths using PK-FK relationships has been extensively investigated by many authors. COCOA[11] proposes a framework in which correlation coefficient between a target variable and column's table plays the most important role in evaluation of the join paths. ARDA[12] incorporates other filter methods such as variance of variables during calculation of the join score. Aurum's Enterprise Knowledge Graph [13] extends the ranking system of the joins by also investigating what new possible joins can be obtained by performing one join.

What one can see as a recurring pattern within the solutions to the problem is a usage of a scoring system between tables before the join. The scoring system is based on characteristics of tables which are deemed to find an optimal features for the training of the ML model. However, the proposed scoring system in the related work is model agnostic (does not take into account on which ML model the data will be trained on). This paper tries to find the heuristics for the joins suitable for the random forest classifier. In the evaluation section it tries to investigate whether those heuristics are applicable to other models. Thus, evaluating whether the assumption of model agnostic data augmentation implicitly used in the related work is suitable.

III. Proposed data augmentation framework

We want to use the best performing characteristics within the proposed framework - partial correlation to judge on whether to join a table or not. At the same time to obtain this characteristics one has to posses knowledge about dependencies between features from both tables. To calculate partial correlation we have to join the tables. But do we need to know this characteristics perfectly? In the proposed data augmentation framework join is sampled by randomly selecting one% of rows from the outer table and joining them with the inner table. After the sample join, the correlation coefficient between features from the inner table and target feature is estimated while controlling for all of other features in the inner table. If the partial correlation exceeds the threshold passed to the framework as a hyper-parameter, non-sample join is performed and data is being augmented. The procedure repeats until no neighbouring table exists for which estimated partial correlation exceeds the given threshold.

Algorithm 1 PCADA routine

```
1:  function PCADA(target_table, threshold)
2:      result ← target_table
3:      frontier ← target_table's neighbours
4:      while frontier is not empty do
5:          current ← pop visited
6:          for all n ∈ current's neighbours do
7:              s ← result sample join current
8:              ave_pc ← CalcPC(s, n)
9:              if ave_pc >= threshold then
10:                 result ← result join current
        return result
```

Estimation of the partial correlation by performing the join only on small sample of the rows takes into account the trade-off between choosing the best characteristic for predicting feature importance and the time required to compute the characteristics (avoiding full join).

On the other hand, the reader has to be aware that limiting the number of rows for the join by a factor of 100, leads to time improvement for the join that is worse than 100 times. Randomly sampling the outer table destroys spatial data locality, which is assumed during buffer management in modern DBMS[14], thus resulting in more cache misses during joins. To get around this problem, one can propose a framework in which the rows are not randomly sampled, but the first one% of rows from the outer table is selected. PCADA does not use this method, as it introduces bias in estimation of the partial correlation (the rows can be ordered by some property, in the worst case by the values of the target column).

IV. Evaluation

The research is divided into two parts. The section IV-A aims at discovering heuristics that makes data augmentation optimal. IV-B focuses on suitability of the found heuristics during joins comparing them to join-all methods and no joins for common data sets.

A. Why partial correlation?

1) Methodology:

a) Measure of optimality: In order to find optimal characteristics of features for joins we have do establish what we define as optimal feature. The most natural measurement of an importance of a feature is an accuracy of the model trained on it.

There is one profound drawback when choosing accuracy as measurement of importance of a feature. Training the model only on the examined feature would not capture all of characteristics of a Random Forest classifier. In fact it would make the model linear. In order to properly judge the importance of a feature based on the accuracy one would have to perform an exhaustive search on all of the columns - training the model on the power set of all of the features and calculating the average accuracy of all of the subsets that contain the given feature. The power set grows exponentially with addition of new feature. Combining this with the fact that random forest is an expensive algorithm on its own, as it requires running decision tree multiple times, makes examining the optimality based on accuracy infeasible for the study.

The above mentioned discovery has been analyzed by the authors of Boruta algorithm in [9]. The authors propose an alternative measure of optimality of a column - Boruta importance. The calculation of Boruta importance is much less computationally expensive than exhaustive search, as it grows linear with added number of columns, instead of exponentially. Considering the above mentioned fact the research uses Boruta importance, as a measure of optimality of a feature.

Nevertheless there exists drawbacks for choosing Boruta importance, as a measure of optimality. Firstly, the said measure is far less interpreatable due to its much more complex calculation method in comparison with mean accuracy. Secondly, the importance in the Boruta algorithm is used to rank the features. Thus, by definition, it produces values that are monotonically, but not necessary linearly linked with optimality of a column. As a result, Spearman correlation coefficient between measured characteristics and feature importance is used, instead of Pearson correlation coefficient.

b) Characteristics examined: Two classes of characteristics can be identified: univariable characteristics and multivariable characteristics. Univariable characteristics examines a column on its own, whereas multivariable characteristics examine a dependency between a column and another columns (usually target feature.)

Within the research, we are limited to finding characteristics that are not very computationally expensive to calculate. Characteristics typical for filter feature selection as described in [6] are mostly used.

For univariable characteristics the following characteristics are investigated within the study:

- variance
- number of missing values
- type of data (categorical vs continuous)
- index of dispersion [15]
- kurtosis
- skewness [16]

For multivariable characteristics the following properties are investigated:

- Gini impurity
- Pearson correlation coefficient between a column and target variable
- Spearman correlation coefficient between a column and target variable
- Partial correlation coefficient [17] between a measured column and target variable while controlling for all other variables
- ANOVA [18]
- Information gain ratio [19]

For each dataset the Spearman's correlation coefficient between characteristics and Boruta importance is calculated.

c) Random forest hyper parameters: Random forest classifier uses many hyper parameters. This include: the number of trees in the forest, function to measure the quality of the split, the maximum depth of the tree, minimum number of samples needed to split internal node, minimum number of samples for a node to become a leaf node, max number of features to consider during examining the best split[20].

As random forest is an expensive algorithm, examining different combinations of hyperparametrs and its dependency between optimal feature characteristics is infeasible. Within the reserach the decision has been made to use hyperparameters as documented in figure 1.

Parameter name	Parameter value
n_estimators	100
criterion	”gini”
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	”sqrt”
max_leaf_nodes	None
min_impiurity_decrease	0.0

Figure 1. Hyper-parameters used in search for optimal feature characteristics

d) Data sets: Characteristics of optimal columns for random forest classifier are data set dependent. Thus, in order to find the general answer to the problem we have to examine the characteristics for various datasets with different properties and average them.

The data sets chosen for the problem are limited to binary and multi-class classification problems. The datasets do not contain regression problems, as those problems are not supported by the Boruta package[9]. The data sets has been chosen based on their popularity on the Kaggle platform. The chosen datasets are:

The proprieties of the data sets can be summarized by the table bellow:

Name	#rows	#columns
Wine Quality	4898	12
Pima Indians Diabetes	768	9
Banknote Dataset	1372	5
Iris Flowers	150	5
Ionosphere Dataset	351	35
Wheat Seeds	210	8
NBA rookie	2217	45
Stroke prediction	5110	12
IBM HR analytics	1470	35
Smart grid stability	60000	14

Figure 2. Properties of the data sets used for the optimal characteristics experiment

The data sets are publicly available at [1].

The chosen data sets are versatile. They include from five input features (Banknote Dataset) up to 35 features (IBM HR analystics). The number of rows range from 150 (Iris Flower) up to 60000 (Smart grid stability). Some data sets contain columns with missing values. The columns consist of categorical and continuous data types. Not all of the data sets are balanced. The variety of data sets will ensure that the found characteristics are generic and data set agnostic.

2) Results: The results of the experiment described above can be summarised by the following figures:

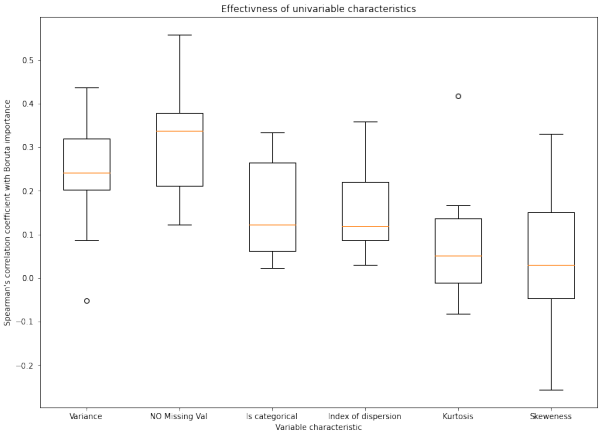

Figure 3. Effectiveness of univariable characteristics measured by Sperman's correlation coefficient between variable's characteristics and Bourta importance

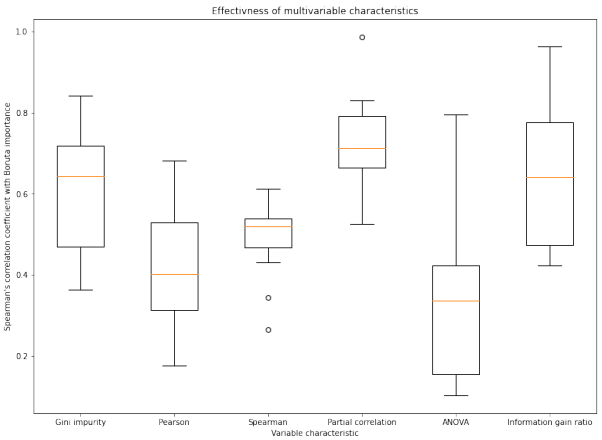

Figure 4. Effectiveness of multivariable characteristics measured by Sperman's correlation coefficient between variable's characteristics and Bourta importance

3) Analysis: By comparing figure 4 and figure 3 one can see that multivariable characteristics are much better in predicting importance of a column than univariable characteristics. This suggests that optimal framework for the data augmentation process should not only take into consideration the table on its own, but also its relation with other tables and target variable.

Characteristics that are better at predicting feature importance are more computationally expensive or require

[1]https://github.com/oskarlorek/pcada_tests

keeping more statistics within the database. Calculating the variance does not require performing any cross table analysis, whereas calculation of Pearson's correlation coefficient between examined feature and target variable requires joining the tables, or marinating the statistics on this characteristics in the database.

Maintenance of simple statistics such as Pearson correlation coefficient can be considered a feasible process. On the other hand, characteristic that performed the best - partial correlation, which measures the correlation between column and target feature, while controlling for other variables requires access to many columns, thus avoiding the join may be infeasible.

B. Comparing to other Data augmentation algorithms

After finding suitable metric for evaluation of the suitability of join - partial correlation and proposing the framework, one has to evaluate its performance and accuracy.

The purpose of this section is to show that:

- augmentation increases the accuracy of the model
- joining all the tables is more time consuming than using PCADA
- PCADA is applicable to other machine learning models

1) Methodology:

a) Data sets: For the purpose of the experiment four data sets has been collected. The data sets are designed for binary classification with decision trees model. The data sets are publicly available [2].

The proprieties of the data sets can be summarized by the table bellow:

Name	#rows	#features	#tables
Football	1182	58	10
Kidney disease	400	31	4
Steel plate fault	1941	51	8
Titanic	891	16	4

Figure 5. Properties of the data sets used for the performance evaluation experiment

b) Setup: To ensure repeatability of the results, the author has decided to run the experiment on cloud hosted environment. The experiment is run on t3.medium instance of AWS-EC2 located in US East (N. Virginia) region.

The accuracy is measured through 5-fold cross validation.

[2]https://github.com/delftdata/auto-data-augmentation/tree/main/other-data/decision-trees-split

c) Metrics measured: Within the experiment we will measure the accuracy of the models trained on the data sets obtained after running the data augmentation algorithms. We will also compare the accuracy of other ML models notably: tree classifiers (XGBoost[21], CART[22]) and one linear classifier - Support Vector Machine[23]. Training the model on other classifiers is needed to judge robustness of PCADA against ML models different than random forest. As a side product of this, we will also gain insight into whether partial correlation is an optimal characteristics for other classifiers.

Apart from it, we will measure the run-time of the data augmentation algorithm combined with the time needed to train the model. Despite the fact that we do not change the code of the ML models, change of the data augmentation algorithm also influences the time needed to run the model, as the dimension of input table changes the amount of data needed to be processed.

d) Data augmentation algorithms examined: To examine effectiveness of PCADA, we need to compare its performance against other ways to integrate the data sets. Within the research we examine two other frameworks.

JoinAll connects all of the tables, based on PK-FK relationships. Note that JoinAll is particularly suitable for the data sets provided, as their relationships forms a Directed Acyclic Graphs. With self referencing relations, or non-DAG relations JoinAll would have to be modified, as it would produce very long join paths.

NoJoin does not perform any joins. The training of the ML model is performed only on the table with the target column.

2) Results:

After performing the experiments the results can be summarized by the figures below:

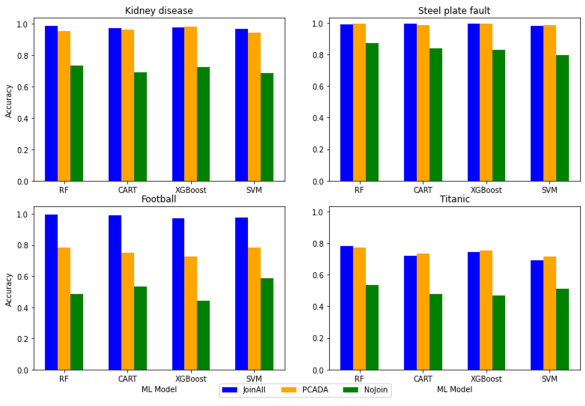

Figure 6. Comparison of accuracy of model trained after running different data augmentation algorithms

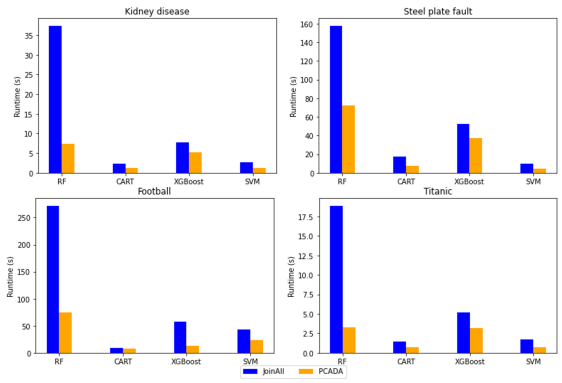

Figure 7. Comparison of run time of PCADA against JoinAll method

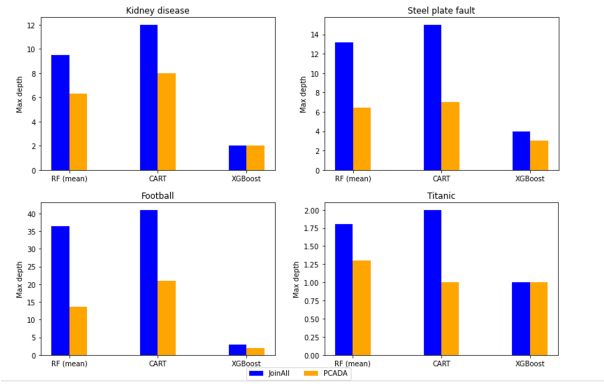

Figure 8. Comparison of (mean) tree depth of PCADA against JoinAll method

C. Analysis

1) Accuracy: By analysing figure 6, we can observe that PCADA achieves much better accuracy than a baseline NoJoin method. Whereas, it achieves similar accuracy to JoinAll for all ML models for all data sets apart from the Titanic deta set. The following result is not surprising, as the number of features obtained through PCADA framework is much larger than with the baseline approach. At the same time, PCADA selects the most important features, thus achieving similar performance to JoinAll aproach.

One can also observe that PCADA achieved better performance than JoinAll approach for the Titanic dataset. This can be explained by over-fitting of all decision tree classifier (CART and XGBoost) and SVM. PCADA eliminates the least significant features, thus guards against over-fitting of the ML model. Note that the accuracy of Random Forrest classifier did not improve while using PCADA, as Random Forrest classifier has already a prevention mechanism for over-fitting - generation of many decision trees and then choosing the most popular vote.

2) Run time: By looking at figure 7, we can observe that PCADA significantly reduces the run-time when comparing it to full data augmentation through JoinAll process. This can be explained by the fact that PCADA performs much less joins and when evaluating the suitability of join it uses non-computationally expenisve process (only sampling one% of join).

Reduction of run-time is especially visible for Random Forest classifier. This is unsurprising since Random Forrest, when given many features, tries to create many decision trees to prevent over fitting. In this case we minimize the time penalty of this process, as PCADA pre-selects only tables with columns that it deems to be interesting.

Note that PCADA's run-time could have been further optimized if we would sample less than one% of joins. However this would result in worse estimation of partial correlation, thus selection of join paths would have been different, in consequence resulting in worse accuracy. This example shows that when designing a data augmentation framework one has to take into account a trade-off between accuracy and runtime.

3) Robustness against different ML algorithms: In the first part of the research we have found that partial correlation is the most suitable metric for measuring importance of features for the random forest classifier. We can see in figures 6, 7, 8 that partial correlation is also a viable heuristic for CART, XGBoost decision trees classifier and Support Vector Machine. This suggests that PCADA can be classified as model agnostic data augmentation framework. Nevertheless, more research has to be performed to find whether partial correlation is the most optimal characteristic for other algorithms.

V. Responsible Research

There are no ethical issues related to the algorithm proposed.

All ideas borrowed from other works are properly cited and included in the references section. To ensure repeatability of the results all experiments were run on cloud environment (t3.medium instance of AWS-EC2 located in US East (N. Virginia) region).

To guarantee reproducibility data sets are publicly available[3], the algorithm is clearly described and the procedure is detailed enough to be repeated by another researcher. All of the hyper parameters used while training the ML algorithms has been described.

VI. Conclusion

Data augmentation framework called PCADA has been developed. It is eagerly evaluating on whether to join a neighbouring based on partial correlation. Partial

[3]https://github.com/oskarlorek/pcada_tests

correlation is estimated using sample join with the neighbouring table.

We have shown in the paper that the most optimal characteristic of a feature for the random forest classifier is partial correlation. Furthermore, we have discovered that multi-variable characteristics perform much better in estimating feature importance than uni-variable characteristics. There exists a trade-off between effectiveness of a characteristic and the time required to compute it.

The work compared PCADA to other baseline data augmentation methods. We have demonstrated that PCADA lies on Pareto frontier when compared it to JoinAll and NoJoin methods. PCADA runs faster than JoinAll, while achieving similar accuracy in many cases. There even exists data sets that overcome JoinAll's performance due to its ability to prevent over fitting, by only joining relevant tables.

Despite the fact that PCADA has been developed with random forest classifier in mind, it was shown that it performs similarly when using other ML algorithms - decision tree classifiers (CART and XGBoost) and linear classifier (SVM). Thus, it suggests that characteristics of optimal features should be considered as ML model agnostic.

VII. Further development

A. Characteristics' performance on other ML models

The paper investigates what is the most optimal characteristic for finding features' importance for random forest classifier. We do not repeat the experiment for other ML models. Despite the fact that using partial correlation for other models achieves similar results, we did not prove that this characteristic is the best. Before generalizing PCADA to other ML algorithms one would have to show that partial correlation is also good at predicting features' importance for them.

B. Determination of optimal sample join ratio

Within PCADA we have proposed to sample one% of join to estimate partial correlation between tables. We have chosen this number, as we deem it as a good trade-off between computation requirements and precision in estimation of the characteristic. However, joining one% of data may be not sufficient to determine partial correlation for small data set. At the same time, for large data sets computation of this heuristic may not be feasible. When using sampling we destroy spatial data locality, thus we reduce the effectiveness of joins. To determine the optimal sample join ratio, one would have to measure the run-time and accuracy of PCADA on large-scale, production-grade data lakes.

C. Non-greedy approach

Currently PCADA only investigates the neighbouring tables to decide on whether to perform the full join operation. In the situations examined this yielded in little reduction of accuracy in comparison with JoinAll approach. However, the implicit assumption that only neighbouring tables contain important features does not always hold. In real-world databases to model many-to-many relationships between entities, an auxiliary table is introduced with PK of both entities. PCADA in its current state would eliminate the possibility of this join, thus reducing possible interesting join paths. To combat this issue, PCADA could not only perform sample join on the neighbouring table, but also on tables that are k-hops away from it similarly to Aurum's Enterprise Knowledge Graph [13].

References

[1] F. Ravat and Y. Zhao, "Data lakes: Trends and perspectives," in *International Conference on Database and Expert Systems Applications*. Springer, 2019, pp. 304–313.

[2] A. Kumar, J. Naughton, and J. 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.

[3] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[4] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[5] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*. Elsevier, 1992, pp. 249–256.

[6] N. Sánchez-Marono, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2007, pp. 178–187.

[7] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering & MIS (ICEMIS)*. IEEE, 2016, pp. 1–5.

[8] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in *Feature extraction*. Springer, 2006, pp. 137–165.

[9] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of statistical software*, vol. 36, pp. 1–13, 2010.

[10] M. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. Molla, "Feature selection for intrusion detection using random forest," *Journal of information security*, vol. 7, no. 3, pp. 129–140, 2016.

[11] M. Esmailoghli, J.-A. Quiané-Ruiz, and Z. Abedjan, "Cocoa: Correlation coefficient-aware data augmentation." in *EDBT*, 2021, pp. 331–336.

[12] N. Chepurko, R. Marcus, E. Zgraggen, R. C. Fernandez, T. Kraska, and D. Karger, "Arda: automatic relational data augmentation for machine learning," *arXiv preprint arXiv:2003.09758*, 2020.

[13] R. C. Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, and M. Stonebraker, "Aurum: A data discovery system," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 1001–1012.

[14] H.-T. Chou and D. J. DeWitt, "An evaluation of buffer management strategies for relational database systems," *Algorithmica*, vol. 1, no. 1, pp. 311–336, 1986.

[15] B. Selby, "The index of dispersion as a test statistic," *Biometrika*, vol. 52, no. 3/4, pp. 627–629, 1965.

[16] M. K. Cain, Z. Zhang, and K.-H. Yuan, "Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation," *Behavior research methods*, vol. 49, no. 5, pp. 1716–1735, 2017.

[17] J. Peng, P. Wang, N. Zhou, and J. Zhu, "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 735–746, 2009.

[18] L. St, S. Wold *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[19] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[20] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[21] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[22] R. J. Lewis, "An introduction to classification and regression tree (cart) analysis," in *Annual meeting of the society for academic emergency medicine in San Francisco, California*, vol. 14. Citeseer, 2000.

[23] S. Yue, P. Li, and P. Hao, "Svm classification: Its contents and challenges," *Applied Mathematics-A Journal of Chinese Universities*, vol. 18, no. 3, pp. 332–342, 2003.