Universidade: presente!

UFRGS PROPESQ

Salão UFRGS 2019 — Conhecimento Formação Inovação

XXXI SIC

21.25. OUTUBRO • CAMPUS DO VALE

A comparative evaluation of aggregation methods for machine learning under vertically partitioned data

Bernardo Trevizan · Mariana Recamonde Mendoza

Introduction

- **Machine learning with vertically partitioned data** happens when features are in distinct sites and due to computational costs or privacy issues cannot be shared. When that happens, state-of-the-art algorithms no longer have a satisfactory performance on local data. **Therefore, aggregation methods can be applied to group local prediction rankings in order to generate a global one.**
- Growing concern about obtaining globally meaningful data mining results without sharing original information among sources have led to different methodologies for vertically partitioned ML. However, **it is still unclear if any of these methods is particularly better than its counterparts**, and whether their performance depends, at least partially, on database's characteristics.

Goals

Perform a comparative evaluation of aggregation methods for vertical data partitioning and investigate their relations to the problem's intrinsic characteristics. Understanding the scenarios in which certain methods are more effective when dealing with classification in vertically partitioned ML, this study should help **build a model to guide the choice of the most promising aggregation methods given the problem's intrinsic characteristics**.

Methodology

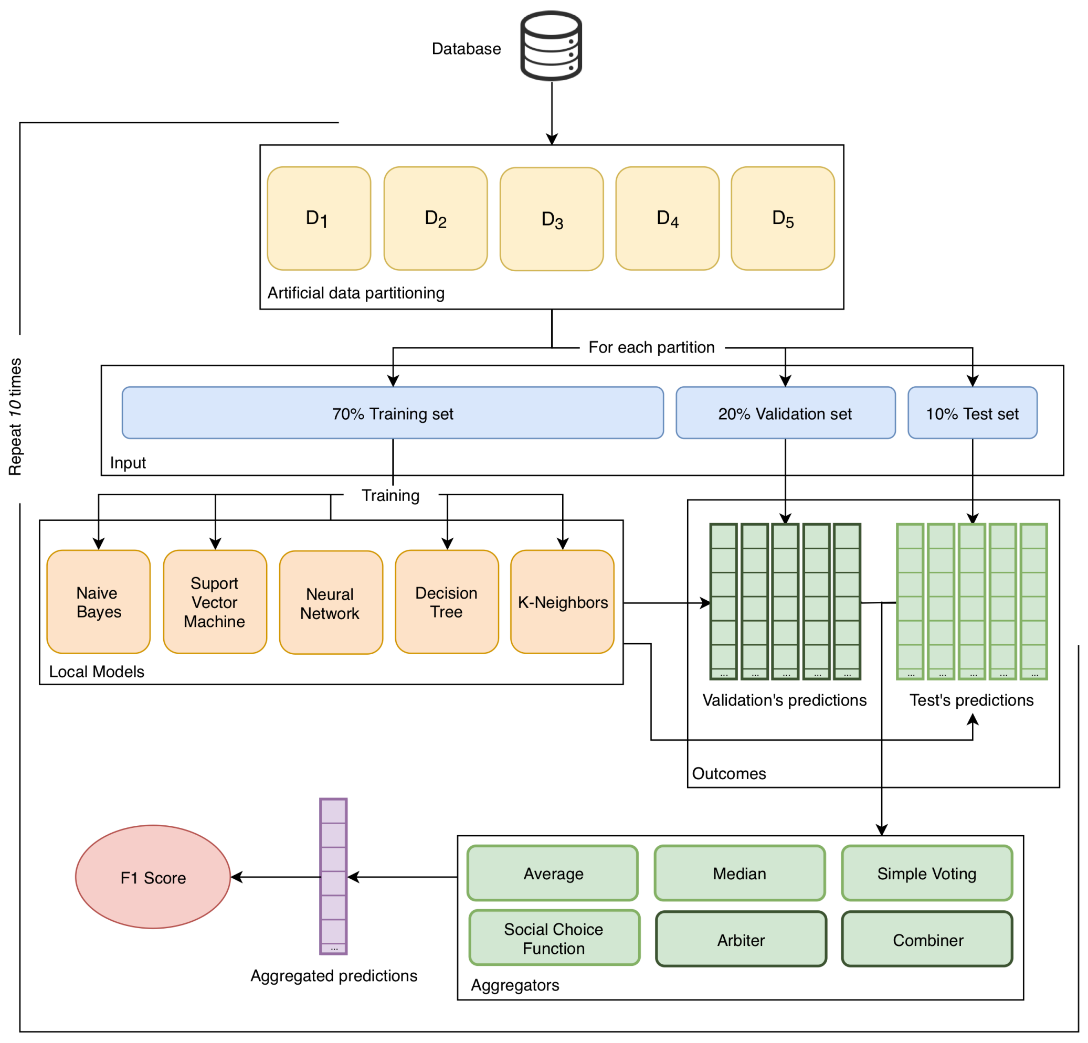
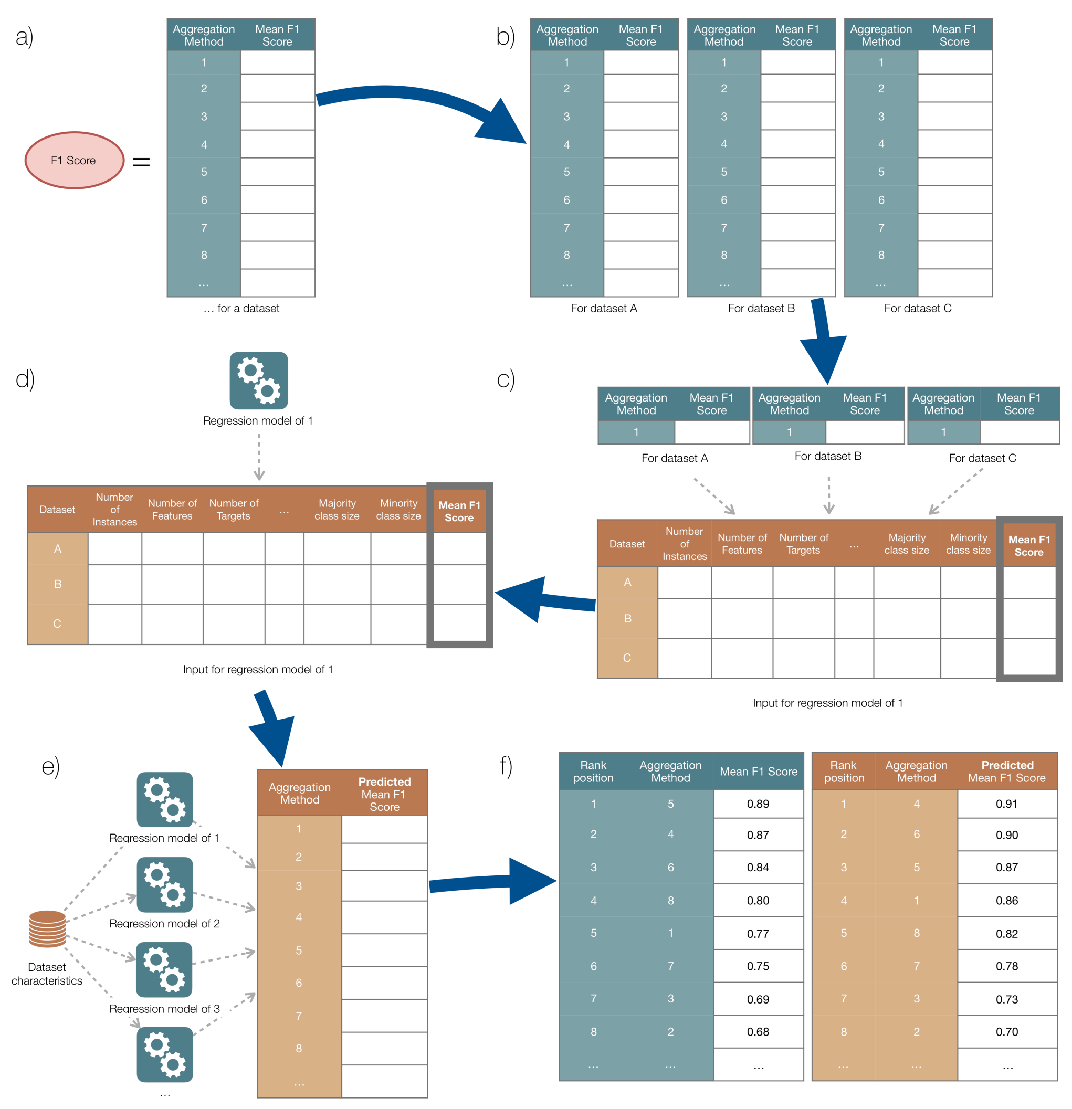


Figure 1. Pipeline for collecting models' performance.



Rank position	Aggregation Method	Mean F1 Score
1	5	0.89
2	4	0.87
3	6	0.84
4	8	0.80
5	1	0.77
6	7	0.75
7	3	0.69
8	2	0.68
…	…	…

Rank position	Aggregation Method	Predicted Mean F1 Score
1	4	0.91
2	6	0.90
3	5	0.87
4	1	0.86
5	8	0.82
6	7	0.78
7	3	0.73
8	2	0.70
…	…	…

Figure 2. Pipeline for data processing.

- **Data Extraction** In order to extract the models' performance the following pipeline (Figure 1) was executed. **Five vertically partitions were artificially generated** from a given database representing five distinct sites. After data partitioning, each partition is assigned to a base classifier, which uses it as input. For training and testing the local and global models, an **adapted 10-fold cross validation** was used in order to minimize any bias in performance evaluation and allow proper comparison among results. This process was repeated ten times, aiming to avoid an evaluation biased by partitions composed solely with the most informative features. **We ran the experiment over 46 databases**, whose main criteria for selection was the diversity in their characteristics, such as (i) number of instances, (ii) number of classes, (iii) number of features, (iv) imbalance degree between classes, (v) average silhouette coefficient, (vi) number of binary features, (vii) majority class size, and (viii) minority class size. The **F1-Score micro-average was used** to evaluate the performance of the models.
- **Data Processing** The F1-Score collected was summarised into its mean (Figure 2a) by dataset (Figure 2b) and grouped by aggregation method (Figure 2c). Each resulting group was joined with the respective datasets' characteristics forming the input for the regression model (Figure 2d). **The regression models trained were used to create a ranking of predicted mean F1-Score** (Figure 2e) for each of the 10 test datasets. The original and predicted rankings were ordered (Figure 2f) and **compared using Kendall tau distance**. Intuitively, the Kendall tau distance measures the number of exchanges needed in a bubble sort to convert one ranking to the other. The concept of **buckets of ranking elements** was also used aiming to reduce the non-significant F1 Score differences. An element of a ranking belongs to a bucket if its F1-Score is in the bucket's interval. Elements in the same bucket have the same ranking position.

Results

We ran the ranking comparison without buckets and due to small prediction errors on F1-Scores the method position on the original and predicted rankings changed. As Kendall tau only considers the elements' position in the ranking, values that are very close to each other may still impact on the distance despite a possibly insignificant difference. In this sense, **we defined buckets' intervals corresponding to the expected performance: [0.00; 0.50) very bad; [0.50; 0.75) bad; [0.75; 0.90) good; [0.90; 1.00] very good.** The results with and without buckets are presented on Table 1.

Dataset	Kendall tau distance (normalised)	Kendall tau distance **using buckets** (normalised)
Sloan digital sky survey	0,54	0,44
Gesture classification	0,52	0,40
Steel plates' fault prediction	0,50	0,41
First order theorem proving	0,59	0,44
Life expectancy prediction	0,47	0,25
Credit approval	0,58	**0,20**
Pulsar star prediction	0,48	**0,47**
Turkey political opinions	**0,64**	0,30
Speech recognition	**0,44**	0,29
Income classification	0,53	0,34

Table 1. Ranking comparison results. The lowest values are highlighted in green and the higher ones in red.

As expected, the **results improved with buckets.** There are cases, such as the Sloan digital sky survey and gesture classification that presented a higher distance between rankings. Thus, in order to understand if the higher distances have been caused by the lower or higher buckets' intervals, we count the disagreements between buckets (Figure 3). The results show that the **middle buckets are disagreeing more than the extreme ones, which suggest that better methods predicted are, in fact, in the first positions on the original ranking.**

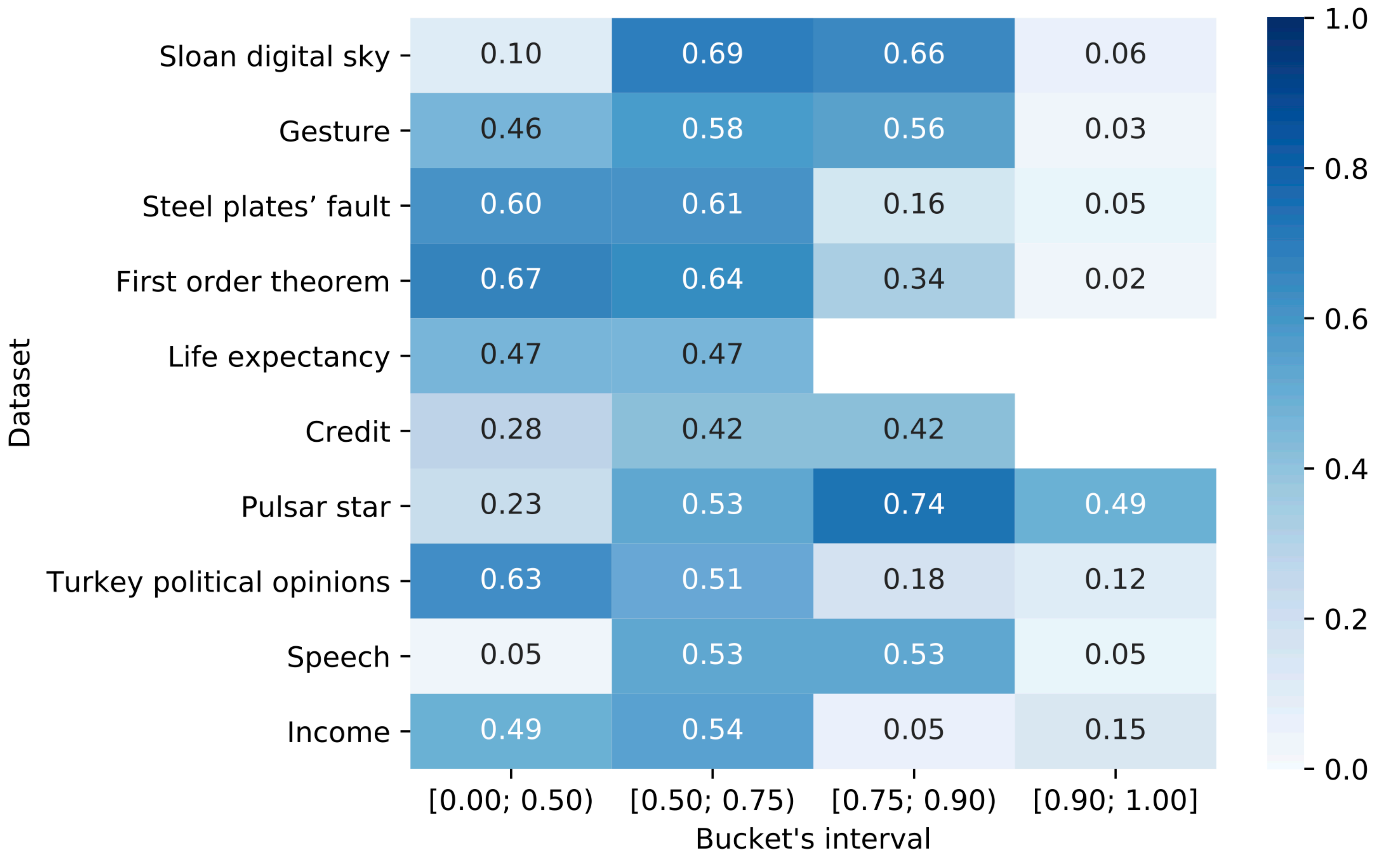


Figure 3. Buckets' disagreements count (normalised). The cells with no value represent the lack of buckets.

Conclusion

- There is no aggregation method that performs better in every case…
- … its performance depends on the problem's intrinsic characteristics as observed on previous work.
- **With this information, we were able to create a model that can be used as practical tool to guide the choice of the aggregation method for vertically partitioned machine learning problems.**
- However, we've ignored one important characteristic due to the difficulty of quantifying it— data pattern — which should explain the results outliers.

Bolsista Bernardo Trevizan
E-mail btrevizan@inf.ufrgs.br

Orientadora Mariana Recamonde Mendoza
E-mail mrmendoza@inf.ufrgs.br

Auxílio Financeiro:
FAPERGS
.inf INSTITUTO DE INFORMÁTICA UFRGS
UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL