Evento	Salão UFRGS 2019: SIC - XXXI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2019
Local	Campus do Vale - UFRGS
Título	A comparative evaluation of aggregation methods for machine learning under vertically partitioned data
Autor	BERNARDO TREVIZAN
Orientador	MARIANA RECAMONDE MENDOZA GUERREIRO

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

Universidade Federal do Rio Grande do Sul

Bernardo Trevizan [1], Mariana Recamonde Mendoza [2]

For centralized data, i.e, data is assembled in a single database, a wide variety of state-of-the-art learning algorithms is available. However, in some situations, features are partitioned on separate databases, which is known as vertical data partitioning. This scenario becomes more frequent as the number of projects producing a large volume of data geographically distributed is increasing rapidly. In such cases, the state-of-the-art ML algorithms may no longer perform satisfactorily, and thus new strategies need to be developed to allow accurate and robust learning under these situations.

Growing concern about obtaining globally meaningful data mining results without sharing original information among sources due to privacy issues and computational costs have led to different methodologies for decentralized ML. In this sense, a variety of aggregation methods are presented in the literature, such as arbiters, combiners and social choice functions, and can be applied to either binary or multiclass problems. 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. Thus, this paper aims to extend the previous works in order to perform a comparative evaluation of aggregation methods for vertical data partitioning and investigate their relations to the problem's intrinsic characteristics. This study should help to understand the scenarios in which certain methods are more effective when dealing with classification in vertically partitioned ML.

The first step in running the experiment is to vertically partition a given database among the base classifiers. 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-Measure micro-average was used to evaluate the performance of the models. Then, the evaluation criteria and the datasets' characteristics was used to create decision paths in order to identify the most influential characteristics in models' performances.

We have investigated if data characteristics may influence or be linked to the classification performance achieved by certain aggregation methods. In addition, we evaluated whether any specific characteristic can be an indication of a better performance and therefore, be used to guide the choice of the aggregation method in problems with vertical partitioning of data. Our results show that no aggregation method can consistently maintain its performance, thus evidencing that performance is influenced by the problems intrinsic characteristics. We identified several characteristics in different contexts that are more likely to have an impact on performance. With this information, we were able to create decision paths that can be used as practical tools to guide the choice of the aggregation method for vertically partitioned ML problems.

[1]Bolsista IC - btrevizan@inf.ufrgs.br

[2]Orientadora - mrmendoza@inf.ufrgs.br