Thesis Overview:

Analysis and design of scalable pre-processing techniques of instances for imbalanced Big Data problems. Applications in humanitarian emergencies situations.

María José Basgall
https://orcid.org/0000-0002-7024-847X
III-LIDI, National University of La Plata, Argentina
DaSCI, University of Granada, Spain
PhD in Computer Science
Advisors: Dr. Marcelo Naiouf and Dr. Alberto Fernández Hilario
{mjbasgall, mnaiouf}@lidi.info.unlp.edu.ar, alberto@decsai.ugr.es

Motivation

The enormous volume of data from different sources, really varied in its typology, generated and processed at great speed, is known as Big Data. The importance of data lies in extracting knowledge from it. Hence, being able to take advantage of a large amount of data allows us to explore and better understand the problems, providing a priori higher quality solutions. To do this, applying Machine Learning for the generation of models is essential, as well as Smart Data so that these models reflect reality and support decision-making. However, it must be noted that the Machine Learning techniques that until now have offered good results are not always able to handle Big Data due to scalability issues. For this reason, they need to be adapted to work in distributed environments, or new techniques or strategies need to be created to deal with this new scenario.

In addition, datasets can usually have certain undesired characteristics or complexities that interfere with the effectiveness of the knowledge extraction process, so they must be preprocessed due to the fact that most learning models assume that the data are free of those characteristics.

Therefore, and since there are few scalable solutions capable of handling Big Data related to this topic, this thesis addresses the distributed and scalable pre-processing of Big Data sets, in order to obtain good quality data, known as Smart Data. Particularly, it focuses on classification problems, and on addressing the following characteristics: (a) imbalanced data; (b) redundancy; (c) high dimensionality; and (d) overlapping.

Objectives

The following specific objectives are established for the aforementioned purpose:

- Enable a state-of-the-art algorithm widely used for the treatment of class imbalance in traditional data scenarios (*Small Data*), to be able to obtain adequate results from large datasets in a distributed manner and in reasonable execution times.
- To design and to implement a fast and scalable methodology for the reduction in both instances and attributes for Big Data sets with high redundancy and dimensionality, while maintaining the predictive capacity of the original dataset.
- To design and to implement a strategy for scalable data characterisation in the context of Big Data classification, focusing on the ambiguous areas of the problem.
- To apply the knowledge acquired during the development phase to solve problems of interest related to humanitarian emergencies.

Contributions

Regarding imbalanced data, a widely used sequential technique of the state-of-the-art in *Small Data* scenarios, is called Synthetic Minority Over-sampling TEchnique (SMOTE). SMOTE is an instance pre-processing technique for balancing class representation by synthetic generation of instances. In this thesis, SMOTE-BD was presented, a SMOTE for Big Data based on a study of the necessary particularities for its design to be fully scalable, and also its behaviour to match as closely as possible the state-of-the-art sequential technique. A variant of SMOTE-BD, called SMOTE-MR, was also introduced, which follows a design that processes data locally at each node. These contributions improve the scalability and computational time required, while maintaining the predictive capability of the subsequent classifier.

In relation to redundancy and high dimensionality, FDR^2-BD was presented, a scalable methodology to reduce a Big Data set in a dual way (reduction of attributes and instances), with the premise of maintaining the predictive quality with respect to the original data. The proposal is based on a cross-validation scheme where a hyperparameterisation process is conducted. FDR^2-BD allows us to know if a given dataset is reducible while maintaining the predictive power of the original data within a threshold. Therefore, our proposal informs which data attributes are the most important ones and what is the percentage of uniform instance reduction that can be performed. The results showed the strength of FDR^2-BD by obtaining very high reduction values for most of the studied datasets, both in terms of dimensionality and proposed instance reduction percentages. We reached up to 70 % feature reduction and 98 % instance reduction, for a maximum accepted predictive loss threshold of 1 %.

Regarding overlapping, GridOverlap-BD was presented, a methodology for scalable characterisation of Big Data classification problems, which relies on grid-based feature space partitioning. GridOverlap-BD allows to identify or characterise problem areas in two typologies: pure and overlapping. In addition, a complexity metric derived from applying GridOverlap-BD was introduced, with focus on quantifying the overlap present in the data. From the experimentation, it was observed that both the characterisation of problem areas and the quantification of the degree of overlap were effectively performed for the datasets used. This implies a pioneering scalable and fully agnostic (model-independent) approach for the characterisation of Big Data problem instances.

Finally, the application of our contributions in a real context and of such relevance as humanitarian emergencies, using public imbalanced datasets was shown. Improvements could be seen by applying SMOTE-BD, and the complexity of each problem could also be understood by obtaining information about the overlap present in the data by applying GridOverlap-BD.

Publications

The main scientific publications that support this thesis are the following:

Basgall, M. J., Naiouf, M., & Fernández, A. (2021). FDR^2-BD: A Fast Data Reduction Recommendation Tool for Tabular Big Data Classification Problems. Electronics, 10(15), 1757.

Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A., & Herrera, F. (2019). An Analysis of Local and Global Solutions to Address Big Data Imbalanced Classification: A Case Study with SMOTE Preprocessing. Cloud Computing and Big Data (Vol. 1050, pp. 75–85). Springer International Publishing.

Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A., & Herrera, F. (2018). SMOTE-BD: An Exact and Scalable Oversampling Method for Imbalanced Classification in Big Data. Journal of Computer Science and Technology, 18(03), e23.

Citation: M.J. Basgall. A *Thesis Overview: Analysis and design of scalable pre-processing techniques of instances for imbalanced Big Data problems. Applications in humanitarian emergencies situations.* Journal of Computer Science & Technology, vol. 22, no. 2, pp. 183-184, 2022.
DOI: 10.24215/16666038.22.e15