ISSN 1870-4069

Graphs to Face the Class Imbalance Problem in Big Data

A. Guzmán-Ponce[1,2], R.M. Valdovinos-Rosas[1], J.S. Sánchez-Garreta[2], J.R. Marcial-Romero[1]

[1] Universidad Autónoma del Estado de México, Facultad de Ingeniería, Mexico

[2] Universitat Jaume I, Castelló de la Plana, Institute of New Imaging Technologies, Department of Computer Languages and Systems, Spain

aguzmanp643@alumno.uaemex.mx, {rvaldovinosr,jrmarcialr}@uaemex.mx,sanchez@uji.es

Abstract. Class imbalance is one of the data complexities widely studied in the field of Data science. The class imbalance problem occurs when one class is strongly over-represented in comparison with the other classes, biasing the learning towards the most represented class. Due to the large volume of data that needs to be processed in Big Data context, is imperative to clean the data for diminish the volume and improve the results. In this way, graph theory is becoming a popular technique in data science to given solutions in real-world problems by transforming them in terms of vertices and edges. In order to face the class imbalance problem, we proposed a graph-based under-sampling method. This method was experimentally validated by using a collection of two-class imbalanced big data sets. The experimental results show a competitive in the classification performance measured by the geometric mean when we compare to several state-of-the-art methods.

Keywords: Big data, graph theory, class imbalance, under-sampling.

1 Introduction

The notion of Big Data is a consequence of the fast generating data, thus brings challenges for the class imbalance problem due to the Big Data characteristics (Volume, Velocity, Variety, Veracity and Value). In this sense, the class imbalance problem in Big Data became a challenging gap to develop strategies in order to give a solution without reducing the classifier performance.

In a binary data set is said to be imbalanced when one of the classes has a lower number of instances than the other, called minority class or positive class (C^+), while the class with a high number of instances is named as majority class or negative class (C^-) [4].

In Big Data context, the most common strategy proposed to deal the class imbalance problem have been the resampling techniques, which are divided into two solve direction: over-sampling consist on increasing the number of instances in the C^+ and the under-sampling, which remove instances from the C^-. This research is focused on the under-sampling approach because recent studies on Big Data [4] have shown that under-sampling methods produce better results than over-sampling methods. Moreover, much data is not necessary, so reducing the data size in Big Data sets became a need.

Fostered by the fast proliferation of the use of graph-theory in data science, where the aim is extracting knowledge from graph topologies, for instance, the clustering communities or feature selection proposed both of them consider weighted graphs [3,6]. Faced with this reality, we introduce the use of graph theory for facing the class imbalance problem in Big Data. The main contribution of this research is: *The use of graph theory to obtain an induced subgraph, which allows getting the borderline of the negative class in Big Data sets to face the class imbalance problem.*

2 Related Work

Most of the existing methods in Big Data have been developed with MapReduce, which is composed of two functions that let parallelize processing data: the Map function split the data into different subsets of data, and the Reduce function which fuse the local outputs into a single final result [5].

Some researches have been conducted through scaling well-know resampling methods [5]. The classic over-sampling method ROS that replicates randomly instances from C^+ until has the same size of C^+, and the classic under-sampling method RUS that removes randomly instances from C^- until has the same size of C^+. Del Río et al. [5] scaled the above algorithms mentioned through MapReduce. In ROS, the Map function randomly balances the positive class by replicate instances, while the Reduce function took all results generated by each Map function and randomly took instances to balance the data set. In the RUS algorithm, the Map function groups all the instances by class, while Reduce collects and balances by removing randomly negative class instances.

In Big Data there are two proposals that use the SMOTE algorithm (Synthetic Minority Oversampling TEchnique) [2]: SMOTE-MR [2] uses the same partition that an instance belongs for computing the k-neighbors from an instance from the positive class, and SMOTE-BD [1] which principal difference consists of computing the k nearest neighbors based on the *kNN-IS*.

3 Hypothesis or Research Objectives

With the use of graph theory to obtain an induced subgraph can be getting the borderline of the negative class in Big Data sets to face the class imbalance problem in terms of performance of a classifier.

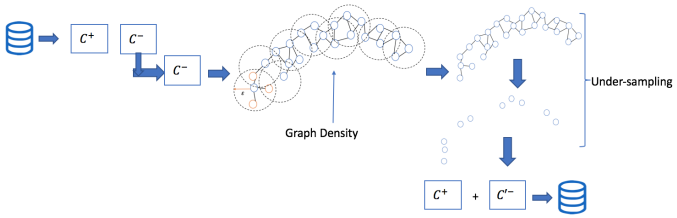

Fig. 1. The steps of Graph-based under-sampling method.

4 Methodology

To achieve the aim of this research, the methodology followed is described as follows:

- Acquisition of data: For carried out the experiments, we use 12 two-class imbalanced Big Data sets taken from the UCI Machine Learning repository, that consider different imbalance ratio (i.e the ratio of the number of instances from the C^- to the number of instances from the C^+). The biggest repository has 4954752 instances and 28 features.
- Data manipulation: In this phase the under-sampling proposal was developed. For that, we obtain an induced subgraph, which allows getting the borderline of the C^- (Figure 1).
  In general, the proposal consists on create a weighted graph based on the density of the data set. Given $v \in C^-$ a vertex, an edge is built according to *eps-neighborhood of an instance* p such that $N_{eps}(p) = \{q \in C^- | dist(p,q) \leq eps\}$, where $dist$ is the Euclidean distance. Thus an edge is built from p to each instance in $N_{eps}(p)$ and their corresponding weighted is given by the dist, this procedure uses MapReduce by using the same partition that an instance belongs.
- Analysis of data: In order to compare the performance behavior of under-sampling proposal on Big Data, we carried out the experimental study with rebalancing methods: RUS, ROS, SMOTE-MR, and SMOTE-BD.

5 State of the Research and Preliminary Results

For each Big Data set, Table 1 reports the geometric mean (averaged across the 5 fold cross-validation) of the Decision Tree classifier from the MLlib Spark API, along with the Friedman's average rank for each algorithm.

According to Friedman's average rank, the best resampling method for the Big Data sets used is ROS, this suggests that duplicate instances extend the sample space, however, this increases the data size. Notwithstanding, as we can see, the top three methods are composed by under-sampling techniques such as RUS and our proposal. Thus, these results show that the proposed method provides competitive advantages with respect to other state-of-the-art methods.

Table 1. mean results obtained per data set, the best performance value is stressed in bold.

Data set	Baseline	ROS	RUS	SMOTE-MR	SMOTE-DB	GraphDensity
poker1	33.7	**53.3**	52.9	42.5	33.7	51.8
SEA	82.0	82.9	82.9	82.9	82.0	**83.0**
Agrawal	94.4	**95.1**	95.0	94.4	94.6	94.8
MiniBooNE	85.2	87.9	88.0	87.9	**88.1**	84.0
Susy	68.8	**76.8**	76.7	76.3	76.3	76.7
Click	16.2	62.1	**62.2**	56.2	56.8	61.9
poker0	17.0	**58.1**	56.2	59.7	53.9	56.2
HEPMASS	71.9	**83.3**	**83.3**	66.8	65.5	83.0
HIGGS	11.6	66.0	65.8	64.4	64.9	**66.1**
Covtype	72.8	**93.2**	**93.2**	92.6	93.0	**93.2**
Credit	87.7	91.9	91.3	**93.0**	92.7	90.0
RLCP	10.3	**93.2**	**93.2**	93.0	93.1	**93.2**
Avg. rank	5.63	**1.92**	2.33	3.96	4.08	3.08

6 Conclusions

In this short paper, a graph-based method to face the class imbalance problem in Big Data is proposed. The obtained results show the potential of graph theory to face this problem. So, the future work is focusing on improving the performance of this proposal through the study of other intrinsic data characteristics. In the same way, an experimental comparison of the behaviour of this and another resampling methods will be addressed for concluding the PhD Thesis.

References

1. Basgall, M., Hasperué, W., Naiouf, M., Fernández, A., Herrera, F.: An Analysis of Local and Global Solutions to Address Big data Imbalanced Classification: A Case Study with SMOTE Preprocessing. In: Naiouf, M., Chichizola, F., Rucci, E. (eds) Cloud Computing and Big Data. pp. 75–85. Springer International Publishing, La Plata, Argentina (2019)
2. Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A., Herrera, F.: SMOTE-BD: An Exact and Scalable Oversampling Method for Imbalanced Classification in Big Data. Journal of Computer Science and Technology, vol. 18, no. 3, pp. 23–28 (2018)
3. Newman, M. E. J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E, vol. 69, pp. 026113 (Feb 2004)
4. Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J., Granda-Gutiérrez, E.: Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem. Applied Sciences, vol. 10, no. 4, pp. 1276 (2020)
5. del Río, S., López, V., Benítez, J. M., Herrera, F.: On the use of MapReduce for imbalanced Big Data using Random Forest. Information Sciences, vol. 285, pp. 112–137 (2014)
6. Zhang, Z., Hancock, E. R.: A graph-based approach to feature selection. In: Jiang, X., Ferrer, M., Torsello, A. (eds) Graph-Based Representations in Pattern Recognition. pp. 205–214. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)