Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 258 (2015) 5–38

FUZZY sets and systems

www.elsevier.com/locate/fss

Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data

Victoria López *, Sara del Río, José Manuel Benítez, Francisco Herrera

Dept. of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, Granada, Spain

Available online 24 February 2014

Abstract

Classification with big data has become one of the latest trends when talking about learning from the available information. The data growth in the last years has rocketed the interest in effectively acquiring knowledge to analyze and predict trends. The variety and veracity that are related to big data introduce a degree of uncertainty that has to be handled in addition to the volume and velocity requirements. This data usually also presents what is known as the problem of classification with imbalanced datasets, a class distribution where the most important concepts to be learned are presented by a negligible number of examples in relation to the number of examples from the other classes. In order to adequately deal with imbalanced big data we propose the Chi-FRBCS-BigDataCS algorithm, a fuzzy rule based classification system that is able to deal with the uncertainly that is introduced in large volumes of data without disregarding the learning in the underrepresented class. The method uses the MapReduce framework to distribute the computational operations of the fuzzy model while it includes cost-sensitive learning techniques in its design to address the imbalance that is present in the data. The good performance of this approach is supported by the experimental analysis that is carried out over twenty-four imbalanced big data cases of study. The results obtained show that the proposal is able to handle these problems obtaining competitive results both in the classification performance of the model and the time needed for the computation.

Keywords: Fuzzy rule based classification systems; Big data; MapReduce; Hadoop; Imbalanced datasets; Cost-sensitive learning

1. Introduction

The development and maturity of the information technologies has enabled an exponential growth on the data that is produced, processed, stored, shared, analyzed and visualized. According to IBM [1], in 2012, every day 1.5 quintillion bytes of data are created, which means that the 90% of the data created in the world has been produced in the last two years. Big data [2] encompass a collection of datasets whose size and complexity challenges the standard database management systems and defies the application of knowledge extraction techniques. This data

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: vlopez@decsai.ugr.es (V. López), srio@decsai.ugr.es (S. del Río), J.M.Benitez@decsai.ugr.es (J.M. Benítez), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.fss.2014.01.015
0165-0114/

comes from a wide range of sources such as sensors, digital pictures and videos, purchase transactions, social media posts, everywhere [3].

This generation and collection of large datasets has further encouraged the analysis and knowledge extraction process with the belief that with more data available, the information that could be derived from it will be more precise. However, the standard algorithms that are used in data mining are not usually able to deal with these huge datasets [4]. In this manner, classification algorithms must be redesigned and adapted considering the solutions that are being used in big data so that they are able to be used under these premises maintaining its predictive capacity.

One of the complications that make difficult the extraction of useful information from datasets is the problem of classification with imbalanced data [5,6]. This problem occurs when the number of instances of one class (positive or minority class) is substantially smaller than the number of instances that belong to the other classes (negative or majority classes). The importance of this problem resides on its prevalence in numerous real-world applications such as telecommunications, finances, medical diagnosis and so on. In this situation, the interest of the learning is focused towards the minority class as it is the class that needs to be correctly identified in these problems [7]. Big data is also affected by this uneven class distribution.

Standard classification algorithms do not usually work appropriately when dealing with imbalanced datasets. The usage of global performance measures for the construction of the model and the search for the maximum generalization capacity induce in algorithms a mechanism that tends to neglect the rules associated with instances of the minority class.

Fuzzy Rule Based Classification Systems (FRBCSs) [8] are effective and accepted tools for pattern recognition and classification. They are able to obtain a good precision while supplying an interpretable model for the end user through the usage of linguistic labels. Furthermore, the FRBCSs can manage uncertainty, ambiguity or vagueness in a very effective way. This trait is especially interesting when dealing with big data, as uncertainty is inherent to this situation. However, when dealing with big data, the information at disposal usually contains a high number of instances and/or features. In this scenario the inductive learning capacity of FRBCSs is affected by the exponential growth of the search space. This growth complicates the learning process and it can lead to scalability problems or complexity problems generating a rule set that is not interpretable [9].

To overcome this situation there have been several approaches that aim to build parallel fuzzy systems [10]. These approaches can distribute the creation of the rule base [11] or the post-processing of the built model, using a parallelization to perform a rule selection [12] or a lateral tuning of the fuzzy labels [13]. Moreover, a fuzzy learning model can be completely redesigned to obtain a parallel approach that decreases the computation time needed [14]. However, these models aim to reduce the wait for a final classification without damaging the performance and are not designed to handle huge volumes of data. In this manner, it is necessary to redesign the FRBCSs accordingly to be able to provide an accurate classification in a small lapse of time from big data.

Numerous solutions have been proposed to deal with imbalanced datasets [7,15]. These solutions are typically organized in two groups: data-level solutions [16,17], which modify the original training set to obtain a more or less balanced class distribution that can be used with any classifier, and algorithm-level solutions, which alter the operations of an algorithm so that the minority class instances have more relevance and are correctly classified. Cost-sensitive solutions [18,19] integrate both approaches as they are focused in reducing the misclassification costs, higher for the instances of the minority class.

The approaches used to tackle big data usually involve some kind of parallelization to efficiently process and analyze all the available data. One of the most popular frameworks for big data, MapReduce [20], organizes the processing in two key operations: a map process that is responsible for dividing the original dataset and processing each chunk of information, and a reduce process that collects the results provided in the previous step and combines those results accordingly including new treatment if necessary. This approach that divides the original dataset in parts can have a strong pernicious effect when dealing with imbalanced datasets as the data intrinsic characteristics impact is amplified. Specifically, the small sample size [21] is induced when the original dataset is shared out and the dataset shift problem [22] may also be encouraged in the process. The addition of these problems reinforce the necessity of properly dealing with imbalanced datasets, not only for the original imbalance that is present in the data but also for the occasioned problems that arise when the partitions are created.

In this paper, we present a FRBCS that is capable of classifying imbalanced big data which has been denoted as Chi-FRBCS-BigDataCS. The method is based on the Chi et al.'s approach [23], a classical FRBCS learning method, which has been modified to deal with imbalanced datasets and big data at the same time. The usage of a FRBCS