Evolutionary Undersampling for Extremely Imbalanced Big Data Classification under Apache Spark

I. Triguero, M. Galar, D. Merino, J. Maillo, H. Bustince, F. Herrera

Abstract— **The classification of datasets with a skewed class distribution is an important problem in data mining. Evolutionary undersampling of the majority class has proved to be a successful approach to tackle this issue. Such a challenging task may become even more difficult when the number of the majority class examples is very big. In this scenario, the use of the evolutionary model becomes unpractical due to the memory and time constrictions. Divide-and-conquer approaches based on the MapReduce paradigm have already been proposed to handle this type of problems by dividing data into multiple subsets. However, in extremely imbalanced cases, these models may suffer from a lack of density from the minority class in the subsets considered. Aiming at addressing this problem, in this contribution we provide a new big data scheme based on the new emerging technology Apache Spark to tackle highly imbalanced datasets. We take advantage of its in-memory operations to diminish the effect of the small sample size. The key point of this proposal lies in the independent management of majority and minority class examples, allowing us to keep a higher number of minority class examples in each subset. In our experiments, we analyze the proposed model with several data sets with up to 17 million instances. The results show the goodness of this evolutionary undersampling model for extremely imbalanced big data classification.**

I. Introduction

In the recent years, the amount of information that can be automatically gathered is inexorably growing in multiple fields such as bioinformatics, social media or physics. Thus, a new class of data mining techniques that can take advantage of this voluminous data to extract valuable knowledge is required. This research topic is being referred under the term: big data [1]. Big data learning poses a significant challenge to the research community because standard data mining models cannot deal with the volume, diversity and complexity that this data brings up [2]. However, the newly arisen cloud platforms and parallelization technologies provide us a perfect environment to tackle this issue.

The MapReduce framework [3], and its open-source implementation in Hadoop [4], were the first alternatives to handle data-intensive applications, which rely on a distributed file system. The development of Hadoop-based data mining techniques has been widely spread [5], [6], because of its fault-tolerant mechanism (recommendable for time-consuming tasks) and its ease of use [7]. Despite its popularity, researchers have encountered multiple limitations in Hadoop MapReduce to develop scalable machine learning tools [8]. Hadoop MapReduce is inefficient for applications that share data across multiple phases of the algorithms behind them, including iterative algorithms or interactive queries. Several platforms have recently emerged to overcome the issues presented by Hadoop MapReduce [9], [10]. Apache Spark [11] highlights as one of the most flexible and powerful engines to perform a fast distributed computing in big data by using in-memory primitives. This platform allows us to load data into memory and query it repeatedly, making it very suitable for algorithms that use data iteratively.

The class imbalance problem is challenging when it appears in data mining tasks such as classification [12]. Focusing on two-class problems, the issue is that the positive instances are usually outnumbered by the negative ones, even though the positive one is usually the class of interest [13]. This problem is presented in a wide number of real-world problems [12]. Furthermore, it comes along with a series of difficulties such as small sample size, overlapping or small disjuncts [14]. In this scenario, one focuses on correctly identifying the positive examples, but affecting the least to the negative class identification. Various solutions have been developed to address this problem, which can be divided into three groups: data sampling, algorithmic modifications and cost-sensitive solutions. These approaches have been successfully combined with ensemble learning algorithms [15].

Evolutionary undersampling (EUS) [16] falls in the category of data sampling strategies, where the aim is to balance the original dataset. In this case, the balancing is done by undersampling, that is, reducing the number of negative class examples. Differently from random undersampling where the focus is to balance the dataset, EUS has a two-fold objective. 1) To create the balanced dataset; 2) To increase the overall performance over both classes of the problem. In order to do so, a supervised balancing procedure is carried out using a genetic algorithm. Once the dataset is balanced, any standard classifier can be used to build a model that should be able to equally distinguish both classes of the problem. This technique is very powerful when dealing with standard imbalanced problems, however, when shifting to a large-scale context it becomes unfeasible since the search space increases exponentially with the number of instances

This work was supported by the Research Projects TIN2011-28488, TIN2013-40765-P, P10-TIC-6858 and P11-TIC-7765. I. Triguero holds a BOF postdoctoral fellowship from the Ghent University.

I. Triguero is with the Department of Internal Medicine of the Ghent University, 9052 Zwijnaarde, Belgium. E-mails: {isaac.triguero@irc.vib-ugent.be

D. Merino, J. Maillo and F. Herrera are with the Department of Computer Science and Artificial Intelligence of the University of Granada, CITIC-UGR, Granada, Spain, 18071. E-mails: dmerino76@gmail.com, {jesusmh, herrera}@decsai.ugr.es

M. Galar and H. Bustince are with the Department of Automatics and Computation, Universidad Pública de Navarra, Campus Arrosadía s/n, 31006 Pamplona, Spain. E-mails: {mikel.galar, bustince}@unavarra.es

978-1-5090-0623-6/16/$31.00