A Review of Informative Data Level Resampling Approaches for Solving Class Imbalanced Problem

Dako Apaleokhai Dickson
Computer Science Department
Federal University of Technology, Minna
Minna, Niger State, Nigeria
dako.pg919379st.futminna.edu.ng

John Kolo Alhassan
Computer Science Department
Federal University of Technology, Minna
Minna, Niger State, Nigeria
jkalhassan@futminna.edu.ng

Solomon Adelowo Adepoju
Computer Science Department
Federal University of Technology, Minna
Minna, Niger State, Nigeria
solo.adepoju@futminna.edu.ng

***Abstract*— In the field of machine learning, Imbalanced learning being one among the most challenging classification problems which is also very common among application dataset. Although, imbalanced approach has received increasing attention over the years due to the necessity of handling real world dataset which are usually skewed in nature, possessing various data difficulty factors. The goal of this work is the review of resampling techniques to identify if data intrinsic characteristics were mostly considered during the design of resampling technique. It went further to categorise the techniques into distance, cluster and evolutionary based method, from the result of said process, also presented the advantages and disadvantages of each category and finally, stating general achievements and drawbacks in resampling approaches. The total search that was conducted for this work, yielded 227 papers published within the last two decades, with emphasis on the last. These articles from imbalanced data domains went through different filtering methods, before been finally reduced to 52. It was presented in this work that distanced based methods have received more attention when compared with cluster based and evolutionary based method, this may be due to its merits, which have been presented in this work. From several previous works, data intrinsic characteristics have been found to be more problematic to learning classifier than imbalanced problem. However, from the findings of this work, it was established that despite the report by publications that data intrinsic characteristics are more harmful than imbalanced nature of data, most existing resampling techniques do not regard data intrinsic characteristic in their design, this may be due to the popular nature and attention drawn by imbalanced problem in publications. However, there are some limiting factors that also need to be resolved generally on all the resampling methods such as: lack of consideration of possible relevant examples in undersampling process, lack of outstanding examples interrelationship and similarities evaluation methods. For future work, a robust resampling technique that will critically consider data difficulty factors when evaluating the region and the examples to oversample and undersample. Resampling techniques should also be evaluated against the different types of difficulty factor so as to ascertain the difficulty type it is best used on to achieve great result.**

Keywords—Machine Learning, Imbalanced data, Preprocessing, Data Level Approaches, Data intrinsic characteristics, Data difficulty factors

I. Introduction

With the advancement of technology and the internet, there have been a copious data generation every day. Therefore, it becomes important to improve the deep understanding of knowledge discovery (KD) and analysis of raw data to enhance decision-making in different industries. An evolution has been done on classification of data through the learning process. This process becomes more complex when the dataset is imbalanced [1]. Among the challenges of supervised machine learning process, one crucial problem is learning from imbalanced data [2]. Imbalanced data is one of the most sensitive problem in data mining and machine learning, which exist in most real-life datasets [3]. Dataset are said to be imbalanced when one or more of its class(es) has a smaller number of examples (minority class) when compared to other class(es) (majority class) in the dataset by a substantial margin; when the dataset consists of two classes or classes above two it is referred to as binary or multiclass respectively. For example, in a sample of 100 patients, the number of patients negative to a particular deadly illness such as covid-19 is 98 and 2 are positive to the illness [4],[5].

When class imbalanced (as in the covid-19 case illustration) is not looked into during classification, learning algorithms or models can be engulfed by the majority class while the minority class tends to be neglected or undiscovered [5]. Knowing that learning task can be complex with class imbalance [6], it is also important to note that the disproportion between class examples is remarkably not solely the main source of potential difficulties [7], [8], [9]. Therefore, it is important that, when considering and processing class examples ratio between minority and majority classes, to also analyse the data complexity and the data intrinsic characteristics such as sub-concepts, small disjuncts, noise, borderline, rare and outlier regions [10], [4], [11]. When these two degrading factors (that is, class imbalanced and data intrinsic characteristics) occurs jointly in a dataset, they severely affect the recognition of the minority class [8].

The objective of this study is the review of resampling techniques to identify if data intrinsic characteristics where mostly considered during the resampling process. It went further to categorise the techniques into distance, cluster and evolutionary based method, from the result of said process, also presented the advantages and disadvantages of each