Confronting of Imbalance Dataset in Diagnosis of Dementia Disorder

Mohammad Reza Kandovani Asl and Farzaneh Rahbari Bonab

ABSTRACT

Manner of encountering imbalance dataset is being constantly recognized as a challenging issue in data mining. Number of errors in recognizing lower range data are constantly being increased. Imbalance dataset refers to the data that, in which relevant samples of a group are more than another group. Dementia disorder is one of the main diseases, which on time diagnosing and controlling it can be vital for life of the patient. As in most cases, number of relevant samples of patients is less than normal individuals, constantly learning machines face problem of imbalance of dataset. The present study has introduced two PRS and PS methods for purpose of solving imbalance dataset problem in regard with performance of classifiers. Dataset from dementia patients, which its characteristics were derived from ADNI, has been applied in order to implement the case.

INTRODUCTION

One of the main challenges in the current age is facing imbalance dataset [11]. Imbalance dataset can be considered as one of the most important problems with different types of medical diagnosis, intrusion detection and fraud detection software in systems, which can cause reduction of accuracy of classifications used in them [10]. For purpose of training high efficiency and accuracy system, there is need for training data with uniform distribution from existing classes. In many cases, the class that is so important in terms of applied range includes fewer samples than the majority class, which the mentioned class is known as imbalance dataset. When the training is based on imbalance dataset, presented classification system would have low efficiency and accuracy [1].

Solutions presented for purpose of solving problems with imbalance dataset are generally based on three theories as follows: methods that the approach in algorithm level helps classifier learning algorithm to conduct learning process towards minority class; methods in data level that can distribute data using sampling data space, so that no change is created in the learning algorithm and it can meet effects resulted from imbalance dataset in pre-processing step [3]. Different methods of sampling have been considered as applied algorithms and classifications. In sampling methods, usually minority class data would be doubled, so that such imbalance can be removed. In sampling methods, omitting data from majority would be considered, so that imbalance data problem can be solved [8].

Learning machines are constantly being conducted toward the class with more number of samples through training from imbalance data. Hence, balancing positive and negative data can be an efficient action for purpose of increasing accuracy of system.

Research objectives:

In regard with classification problems, number of samples in each class can be varied. Problem of imbalance would be occurred when the number of samples in a class is more than samples in another class and such imbalance can cause inefficiency of classifiers. The main objective of the present study is using hybrid algorithms based on previous balancing methods and support vector for purpose of solving problem of reduction of classification while facing imbalance dataset.
Research literature:

Over the years, many different studies have been conducted on presented methods and better results than before are being observed as time goes. Currently, studies are being conducted using 5 different sampling algorithm methods, which some new changes and improvement of results are being observed using k-nearest neighbor algorithm classifier [5]. Some scholars believe that using classifications based on fuzzy logic can be one efficient method for fighting against imbalance dataset; meaning that size of each class would be obtained based on the frequency and additional samples would be omitted with the adjustment of class. Obtained results indicate effect of the classifiers on solving problem of imbalance dataset [2]. In addition, it has been prove din many cases that c4.5 classifiers and support vector have indicated better performance than others because of high flexibility in regard with encountering imbalance dataset [5].

MATERIALS AND METHODS

Proposed method for purpose of solving problem of imbalance dataset:

Imbalance dataset problem is one of the main problems with learning machines. For purpose of encountering the problem, hybrid algorithms based on support vectors have been proposed for purpose of solving the problem and accuracy of the classifier has been also increased. Proposed methods have been presented in form of 2 different algorithms, which would be analyzed following.

PRS method:

Proposed algorithm presented in this study is based on sampling methods and omission of samples of majority class. As it was mentioned, selected samples through PS method are imbalance. In this method, RS algorithm would be applied for purpose of solving imbalance dataset problem, which can provide better results than the previous methods.

PS method:

The method is the first proposed method in this study, in which samples would be selected with more accuracy and reliability and not only generally, but also some samples would be selected from majority class using Support Vector Machine. The samples, along with all samples in minority class, would be given to the classifier for purpose of training. Investigation of results indicates that problem of imbalance dataset has been solved, but not completely. In regard with the balancing issue, it should be mentioned that the lower the difference between class samples is, the higher efficiency and accuracy of the classifiers would be.

![Diagram](Fig. 1: Schema of the model among statistical parameters of RVM algorithm.)

Methodology:

For purpose of system evaluation, a database including 51 samples with 1775 characteristics related to the dementia disorder has been applied. In order to increase the number of samples, each sample would be combined with 10 samples, which are selected randomly from the class of the same sample, and then would be added to the database. Hence, number of samples in dataset would be increased. From existing samples in database, 154 samples associated with normal individuals were considered as majority class and 77 samples from patients were selected as minority class randomly. From selected samples, 20% of them (45) with equal number in two groups were considered as experimental samples and 80% of them (185) were selected as training samples from both classes in imbalance form. For purpose of evaluating proposed methods named PS and PRS, MATLAB software has been applied.
Results:

Classification by RVM:

RVM algorithm is one of the first and the most important steps in modeling based on linear combination of base functions, which can provide a probable structure for judgment on data [4].

As the present study has investigated support vector machine (SVM), Relationship Vector Machine (RVM) based on performance on basis of probabilities can be a novel method for purpose of encountering problem of imbalance dataset.

Figure 2 has illustrated obtained results from evaluation before and after data balancing through different balancing methods using RVM classifier.

![Graphs showing accuracy, F-measurement, and G-Means evaluation criteria after data balancing through different methods using RVM.]

Fig. 2: Obtained results of evaluation using RVM.

Classification using Naive Bayes classifier:

Figure 3 has illustrated obtained results from evaluation before and after balancing data using different data balancing methods using NB Classifier.

Discussion:

Recently, problem of imbalance of classes has gained attention of scholars in field of data mining. In numerous cases, a class that is so important in terms of application range includes lower modes than the class that is majority class. The class can be named as imbalance dataset. Traditional approaches of data mining have not desirable ability for predicting minority samples, which are under consideration. Unfortunately, in most cases the actual data has such characteristic. For example, in regard with diagnosing rare diseases, network invasions and text mining, data distribution is usually imbalance. At the present study, some methods have been proposed based on data clustering and support vector machine for purpose of classifying imbalance dataset.
Conclusion:

PS and PRS methods have been proposed for purpose of solving problem of imbalance dataset. Evaluations have been conducted on a database of dementia disorders. Investigations have been done on two different classifiers of NB and RVM. Obtained results have been also presented in form of bar graphs in section of results.

ACKNOWLEDGEMENT

We acknowledge from Dr. Alireza Nourazarian for his cooperation.

REFERENCES


