Progress in Nuclear Energy 100 (2017) 355–364

Contents lists available at ScienceDirect

Progress in Nuclear Energy

journal homepage: www.elsevier.com/locate/pnucene

Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors

Manas Ranjan Prusty [a, b, *], T. Jayanthi [a], K. Velusamy [c]

[a] Electronics and Instrumentation Group, Indira Gandhi Centre for Atomic Research, Kalpakkam, India
[b] Department of Computer Science Engineering, ITER, Siksha 'O' Anusandhan University, Bhubaneswar, India
[c] Reactor Design Group, Indira Gandhi Centre for Atomic Research and Homi Bhabha National Institute, Kalpakkam, India

ARTICLE INFO

Article history:
Received 22 December 2016
Received in revised form
24 May 2017
Accepted 19 July 2017

Keywords:
Imbalanced data set
Over-sampling
SMOTE
Sodium cooled fast reactor

ABSTRACT

Traditionally, the plight of imbalanced dataset and its classification quandary has been counteracted mostly using under-sampling, over-sampling or ensemble sampling methods. Among these algorithms, Synthetic Minority Over-sampling Technique (SMOTE) which belongs to oversampling method has had lot of admiration and extensive range of practical applications. SMOTE algorithm works on the principle of oversampling of minority data samples by generating synthetic data. The oversampling happens with respect to each minority sample and eventually it leads to oversampling of the minority data set. In this paper, SMOTE has been modified to Weighted-SMOTE (WSMOTE) where oversampling of each minority data sample is carried out based on the weight assigned to it. These weights are determined by using the Euclidean distance of a particular minority data sample with respect to all the remaining minority data samples. Each minority data sample need not generate equal number of synthetic data in WSMOTE as in the case of SMOTE. The performances of the classifiers based on SMOTE and WSMOTE are compared using few real datasets and eventually tested on events in a sodium cooled fast reactor. Recall and F-measure from the confusion matrix have been identified as the principal metrics to evaluate the performance of the classifier. It is seen that WSMOTE performs better than SMOTE algorithm.

1. Introduction

An imbalanced dataset with two-classes consists of data samples with a huge difference between the number of minority data samples and the majority data samples. The minority dataset consists of the samples of a particular class those are low in number whereas the majority dataset consists of the samples of the other class those are comparatively large in number. Such kind of imbalance in dataset is known as between-class imbalance dataset (Japkowicz and Stephen, 2002) compared to within-class imbalance (Japkowicz, 2001). The performance of the classifier network for such imbalanced dataset is always biased towards the majority dataset because of large number of samples it contains. Hence, the classifier does not classify the minority data samples accurately and more often than not these samples are misclassified. This leads to a big challenge in cases where classifying the minority data samples is of utmost priority compared to the majority data samples. Hence, the necessity of improved performance of a classifier network in classifying minority data samples in an imbalanced dataset has brought in a lot of interest amongst researchers and users. An example of such scenario is classifying the occurrence of a malignant disease among a group of people who have symptoms of that disease. In such a case, only a few people will actually have a malignant disease compared to all. It can be really catastrophic when a true malignant disease sample which in this case in the minority data sample is misclassified. The class imbalance problem is generally encountered in the diagnosis fields such as medical diagnosis (Nahar et al., 2012; Sun et al., 2013), fraud detection (Dal Pozzolo et al., 2014; Fawcett and Provost, 1997), intrusion detection (Chairi et al., 2012; Cieslak et al., 2006), bioinformatics (Yu et al., 2013), data gravitation (Peng et al., 2014), finance risk management (Brown and Mues, 2012) and event identification in nuclear power plants.

A path to counteract such situation is by preprocessing the datasets prior to feeding it as input to the classifier network. The commonly used preprocessing methods for such kind of issue are over-sampling, under-sampling and ensemble learning. A wide

* Corresponding author. Electronics and Instumentation Group, Indira Gandhi Centre for Atomic Research, Kalpakkam, India.
E-mail address: manas.iter144@gmail.com (M.R. Prusty).

http://dx.doi.org/10.1016/j.pnucene.2017.07.015
0149-1970/

range of survey of all the preprocessing methods have been carried out by many researchers (Chawla, 2005; He and Garcia, 2009; Japkowicz and Stephen, 2002). In this paper, the oversampling method is mostly concentrated upon. A widely used oversampling method which is being widely used in many practical applications is the SMOTE method (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002). A series of improvement to SMOTE has been carried out by many researchers from the time it was introduced (Bunkhumpornpat et al., 2009; Chawla et al., 2003; Gao et al., 2011; Han et al., 2005; He et al., 2008; Li et al., 2011; Zeng and Gao, 2009; Zhai et al., 2011). The multiple re-sampling method is an additional approach to tackle imbalanced dataset (Estabrooks et al., 2004).

In most of the SMOTE related oversampling, the amount of oversampling done for each minority data sample is fixed to the oversampling percentage. This means, if the oversampling percentage is 200%, then each minority data sample generates two synthetic data. This approach at the end produces 200% of the whole minority dataset. In this paper, the oversampling for each minority sample is different but eventually it leads to the assigned oversampling percentage. This means, for 200% oversampling of the minority dataset, the amount of generation of synthetic data sample for each minority data sample varies individually but in the end the total amount of oversampling increases by 200% of the initial count of minority dataset. This approach is processed by assigning weights to each of the minority data sample based on its Euclidean distance from rest of the minority data samples. The closer a particular minority data sample from the other entire minority samples, i.e., the shorter the Euclidean distance, the larger is the generation of synthetic data for that particular minority data sample. This modified method is named as Weighted-SMOTE (WSMOTE) as weights are assigned to each minority data sample for the generation of a particular number of synthetic data. In this paper, WSMOTE based classifier performance is investigated and compared with SMOTE based classifier. Some of the real world datasets along with datasets from a sodium cooled fast reactor (SFR) are used for the analysis. A ten-fold cross validation approach is undertaken and the final performance is averaged out of these ten folds.

The rest of the paper is organized as follows. Initially, section 2 explains the SMOTE algorithm briefly followed by section 3 which explains WSMOTE algorithm in detail. Section 4 explains the various performance measures which are generally used to calculate the performance of any classifier using confusion matrix. Further, in section 5, the overall experiment and procedure of approach are explained using the real world datasets. A comparison in performances of classifiers based on SMOTE and WSMOTE is performed in section 6. Section 7 elucidates the performances of these classifiers in SFR dataset in event classification. Finally, the paper concludes in section 8.

2. Synthetic minority over-sampling technique (SMOTE)

SMOTE algorithm was proposed to counteract the imbalanced dataset problem for classification (Chawla et al., 2002). It synthesizes new instances of the minority class by operating in the “feature space” rather than in the “data space”. This is an oversampling algorithm in which each minority data generates *N*% of synthetic data. The percentage increase in the minority data should be in such a way that it is comparable with the number of majority data. This increase in instances of the minority data expands the decision reasons for it in the classifiers.

In this algorithm, some parameters such as *T*, *N*% and *k* is initialized at the beginning where *T* refers to the number of minority class samples, *N*% refer to the percentage of oversampling to be done and *k* denotes the *k* value of the *k* nearest neighbor of a particular minority class sample. The steps involved in generating the synthetic samples are as follows.

Step 1. After this initialization, a minority class sample is chosen whose synthetic data has to be generated.

Step 2. Then, one among the *k* nearest minority class neighbors of that sample is randomly selected.

Step 3. As it is known that a sample consists of number of feature data, a synthetic sample is produced by generating synthetic data for each feature data. A synthetic data is generated by adding a factor to the initial feature data. This factor is calculated in two steps. First, the selected feature data is subtracted from the initial minority feature data. Secondly, this subtracted value is multiplied with any value between 0 and 1.

Step 4. This process is carried out for all the other feature data of a particular minority class sample which generates a row of synthetic sample for that minority class sample.

Step 5. For *N*% oversampling, this process is carried out for a rounded value of (*N*/100) to its nearest integer. This generates *N*% of oversampling of a single minority class sample.

Step 6. This procedure is carried out for all the *T* minority class samples which finally results in *N*% oversampling of all the minority class samples.

3. Weighted SMOTE (WSMOTE)

The WSMOTE method is an oversampling method which assigns weights that decide the number of new synthetic data which needs to be generated using SMOTE for an individual minority data sample. This is a modification to the SMOTE algorithm where each of the minority data generates equal number of synthetic data. The WSMOTE method uses the Euclidean distance of each minority data sample with respect to all the other minority data samples in order to produce a weight matrix as shown is Fig. 1. This weight matrix along with the total percentage of synthetic data generation produces the SMOTE generation matrix using Eq. (1). This ultimately gives the number of synthetic data which needs to be generated for a specific minority data sample.

$$[SMOTE\ Generation\ Matrix]_{T\times 1} = \frac{N \times T}{100}\,[Weight\ Matrix]_{T\times 1} \tag{1}$$

3.1. Steps involved in WSMOTE

1. The minority training dataset is considered which contains *T* number of samples and each sample with *C* number of features. The Euclidean distance (ED) of each of the *T* minority data samples are calculated with respect to all the other minority data as given by Eq. (2). In this equation, $ED_i(m_i,m_j)$ represents the Euclidean distance of the *ith* and *jth* samples and *k* represents the *kth* attribute of that particular sample.

$$ED_i(m_i, m_j) = \sqrt{\sum_{k=1}^{C}\left(m_{i,k} - m_{j,k}\right)^2} \tag{2}$$

Here, $i = [1, 2, \ldots, T]$ and $j = [1, 2, \ldots, T]$ and $j \neq i$. The sum of each of these EDs for each *jth* minority sample gives the ED_i. The ED for all the minority data are calculated and stored in a column matrix