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

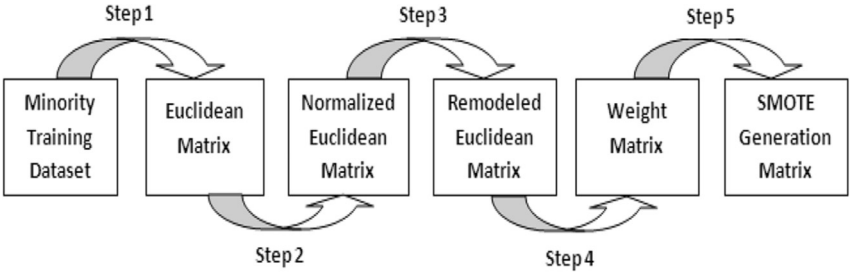

Fig. 1. Block diagram of WSMOTE algorithm.

$ED = [ED_1, ED_2, \ldots, ED_T]'$.

2. This ED matrix is then normalized using the maximum of the ED (ED_{max}) and the minimum of the ED (ED_{min}) and termed as normalized ED matrix (NED) as given by Eq. (3). Normalization is done to map the numbers within a range 0 and 1.

$$NED_i = \frac{ED_{i-}\ ED_{min}}{ED_{max-}\ ED_{min}} \tag{3}$$

3. The NED matrix is then modified to a remodeled normalized ED matrix (RNED). RNED matrix depicts that the lesser the ED of a minority data, the more share it gets to generate the synthetic data out of the total percentage of synthetic data (N%) that needs to be generated. RNED matrix is calculated by subtracting the normalized ED value of each minority data from the sum of all the normalized ED values as represented by Eq. (4).

$$[RNED]_{T\times 1} = sum(NED) - [NED]_{T\times 1} \tag{4}$$

4. Finally, the weight matrix is calculated by finding each minority data share fraction with respect to the total sum of the shares in the RNED matrix as given by Eq. (5).

$$[Weight\ Matrix]_{T\times 1} = \frac{[RNED]_{T\times 1}}{sum(RNED)} \tag{5}$$

5. This weight matrix is used to find the SMOTE generation matrix using Eq. (1).

This is explained with a simple example with T = 5, N = 500% with the ED calculated as in Table 1. The imbalanced dataset considered for this observation is totally fictional. This is just to give a clearer picture on the procedure mentioned above. The ED calculated is the sum of the Euclidean distance of each minority sample with respect to all other minority samples. For example, ED_1 is the sum(ED_{12}, ED_{13}, ED_{14}, ED_{15}) which equals 2 as given in Table 1. These values are again random numbers taken in order to show the working of this algorithm. Similar is the case for calculating the ED of each minority sample.

Table 1
Oversampling of a random dataset using Weighted SMOTE algorithm.

T	1	2	3	4	5
ED	2	1	5	3	4
NED	0.25	0	1	0.5	0.75
RNED	2.25	2.5	1.5	2	1.75
Weights	0.225	0.25	0.15	0.2	0.175
# SMOTE Generation	6	6	4	5	4

Table 1 shows that each minority data sample generates different number of synthetic minority data ranging from 4 to 6 instead of every minority data sample generating 5 synthetic data. This also shows that the smaller the ED, the larger share of synthetic data generation is assigned for that particular minority data sample. As the ED does not depict proper results for high dimensional data, this approach may have to be tuned for handling such datasets. In such kind of situations, a distance metric such as Gaussian based functions which ensembles for high dimensional data might be well suited. A classifier along with a learning algorithm must be independent of the characteristics of the datasets.

4. Performance measures

A confusion matrix is a table which helps to measure the performance of a supervised classifier where the classifier is trained based on known classes. Fig. 2 shows the representation of a confusion matrix for a binary classification problem. True Negatives (TN) are the number of negative examples which are correctly classified as negative. False Positives (FP) are the number of negative examples which are incorrectly classified as positive. True Positives (TP) are the number of positive examples which are correctly classified as positive. False Negatives (FN) are the number of positive examples which are incorrectly classified as negative. In this paper, the majority class samples are the negative class samples and the minority class samples are the positive class samples.

The most commonly used performance measures derived from the confusion matrix are accuracy and error rate. The former is the ratio of the number of all the correctly classified samples to the total number of test samples where as the later is the ratio of all the incorrectly classified samples to the total number of test samples as shown in Eqs. (6) and (7) respectively. It may be highlighted that accuracy and error rate sum up to 1.

$$Accuracy = (TN + TP)/(TN + FP + FN + TP) \tag{6}$$

$$Error\ rate = (FP + FN)/(TN + FP + FN + TP) \tag{7}$$

Accuracy and error rate measures are very deceptive as these are data dependent. In case of imbalanced dataset where the number of

	Predicted Negative	Predicted Positive
Target Negative	TN	FP
Target Positive	FN	TP

Fig. 2. Confusion matrix.

Table 2
Dataset distribution.

Dataset	Minority class	Majority class	#Minority class samples (T)	#Majority class samples (M)	Ratio (T/M)
Ecoli	imU	Remainder	35	301	1:9
Abalone	9	18	42	689	1:16
Wine Quality	8	Remainder	175	4723	1:27
Yeast	ME2	Remainder	51	1433	1:28
Mammography	Calcifications	Non calcifications	260	10 923	1:42

majority samples is too large compared to the minority samples, the classifier gets biased towards the majority samples. In such cases, the accuracy metric results in a very high value and the error rate being very low. These results project as if the classifier is an ideal one which actually is not the real scenario. In such cases, the minority class does not get properly classified yet these two metrics suggest the classifier to be an efficient one. In order to overcome such imbalanced dataset scenarios, there are some other evaluation metrics which state the actual performance of the classifier. These metrics are precision, recall, specificity, fall-out, F-measure and G-mean. These measures are defined as follows.

$$Precision = TP/(TP + FP) \tag{8}$$

$$Recall\ or\ Sensitivity\ or\ True\ Positive\ Rate = TP/(TP + FN) \tag{9}$$

$$Specificity\ or\ True\ Negative\ Rate = TN/(TN + FP) \tag{10}$$

$$Fall{-}out\ or\ False\ Positive\ Rate\ or\ (1 - Specificity) = FP/(TN + FP) \tag{11}$$

$$F{-}Measure = \frac{\left((1+\beta)^2 \times Recall \times Precision\right)}{\left(\beta^2 \times Recall\right) + Precision} \tag{12}$$

$$G{-}Mean = \sqrt{Precision \times Recall} \tag{13}$$

In Eq. (12), β is the co-efficient to adjust the relative importance of precision and recall. Usually precision and recall have equal importance and hence $\beta = 1$. F-measure is the harmonic mean and G-mean is the geometric mean of precision and recall. So, Eq. (12) can be simplified as follows.

$$F{-}Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{14}$$

Precision is the measure of the exactness of the samples which were correctly classified positive out of the samples which are classified positive as shown in Eq. (8). Recall is the measure of the samples which were correctly classified positive out of the samples which were actually positive as shown in Eq. (9). Recall is also known as sensitivity or true positive rate. The importance and information of both precision and recall can be combined together as a measure known as F-measure. F-measure is the harmonic mean where as G-mean is the geometric mean of precision and recall as shown in Eqs. (12) and (13) respectively. Specificity is the measure of the samples which were correctly classified negative out of the samples which were actually negative as shown in Eq. (10). This is similar to sensitivity. It is also called true negative rate. Fall-out is the measure of the samples which were incorrectly classified positive out of all the samples which were actually negative as shown in Eq. (11). Fall-out is also called false positive rate or 1-specificity. Sensitivity and 1-specificity are also used to plot the receiver operating characteristics (ROC) and calculate the area under the ROC curve (AUC) to find the performance of a classifier (Bradley, 1997; Fawcett, 2006; Hanley and McNeil, 1982).

5. Experiment and procedure

5.1. Experiment

In this paper, some real world datasets are considered for analysis purpose as shown in Table 2. All these datasets are converted to binary dataset with minority to majority class sample ratio varying from 1:9 to 1:42. All these datasets are available in the University of California at Irvine (UCI) Repository.[1] The samples which have any missing attributes are deleted from the dataset.

5.1.1. Datasets

1. E coli dataset: This dataset contains the data regarding the protein localization sites for *Escherichia coli*. This dataset contains 8 attributes along with the class name. The dataset is segregated into a total of 8 classes with each class denoting one of the 8 protein localization sites. As this paper is concentrated on binary imbalanced dataset problems, inner membrane cleavable signal sequence (imU) is considered as the minority class where as all the others are considered as majority class. Finally, the minority to majority class sample ratio is 35:301 which is approximately 1:9.
2. Abalone dataset: This dataset is used to find the age of the abalone based on 4177 samples and each sample consisting of seven features. The age of the abalone varies from 1 to 29. Hence, the total number of classes in this dataset is 29. In this paper, class 9 is considered as the minority class with 42 samples and the class 18 is considered as the majority class with 689 samples. The ratio of minority to majority class sample in this dataset is 1:16.
3. Wine Quality dataset: This dataset classifies the white wine quality which ranges from 0 to 10 containing 4898 samples with each sample having 11 attributes. In this paper, class 8 is considered as minority class with 175 samples and the rest 4723 samples as majority class. The ratio of minority to majority class sample in this dataset is 1:27.
4. Yeast dataset: This dataset classifies the localization position of protein in yeast with 1484 samples and each sample consisting of 8 attributes. Each sample can be classified to any one of the 10 localization site. In this paper, ME2 (membrane protein, uncleaved signal) is treated as minority class sample with 51 samples and the rest 1433 samples as majority class. The ratio of minority to majority class sample in this dataset is 1:28.
5. Mammography dataset: This dataset has 260 calcifications out of 11 183 samples. These samples have 6 attributes in each sample. It is important that most of the 260 samples should be

[1] http://www.ics.uci.edu/~mlearn/.

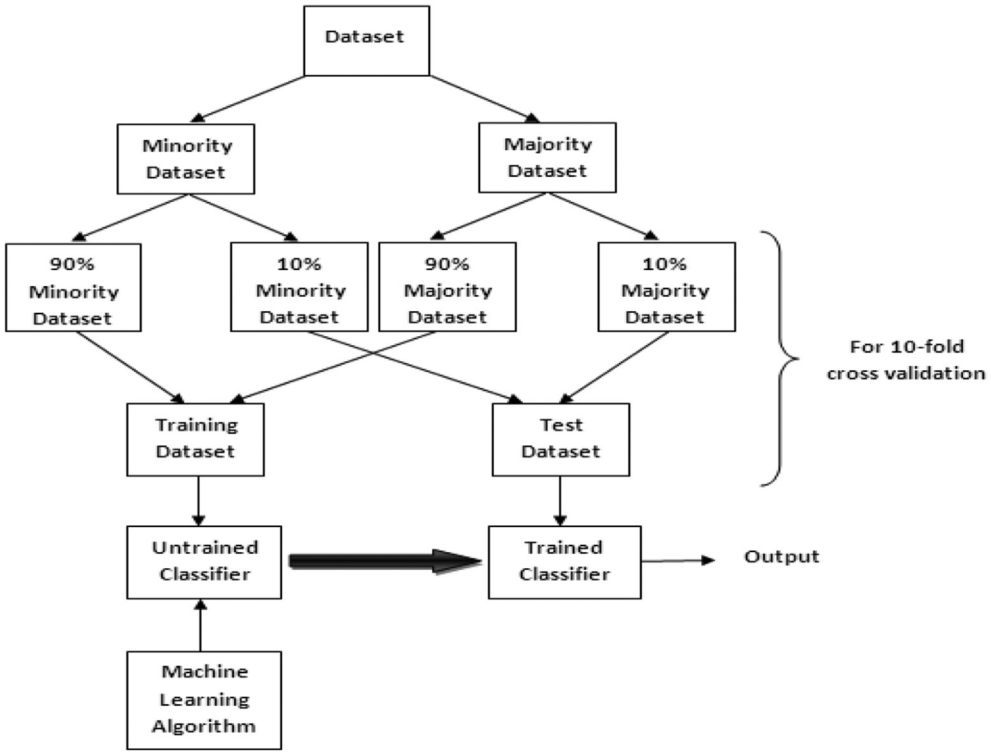

Fig. 3. Preprocessing procedure of a dataset.

classified correctly. Hence, these 260 calcifications are considered as minority class samples where as the rest 10 923 samples are considered as majority class samples. The ratio of minority to majority class sample in this dataset is 1:42.

5.1.2. Classifier network

The classifier network used in this paper is a multi layer perceptron (MLP) network which is trained by back propagation algorithm using Levenberg-Marquardt optimization. The number of nodes in each hidden layer was equal to the number of attributes present in their respective datasets.

5.2. Procedure

As shown in Fig. 3, in this paper, each dataset followed a specific procedure before feeding as a classifier input along with the machine learning algorithm. Each dataset is initially segregated into minority dataset and majority dataset based on the class label of each sample in the dataset. After this step, the minority dataset is divided into two sets, one with 10% and other with 90% of the minority dataset. In the same way, the majority dataset is also divided into two sets, one with 10% and the other with 90% of the majority dataset. This 10% of minority dataset and 10% of majority dataset are clubbed together to form the test dataset. The other 90% of minority dataset and 90% of majority dataset were clubbed to form the training dataset. Now this training dataset which contains both majority and minority dataset is fed as input to the classifier. This classifier is made to learn using any machine learning algorithm. The classifier is basically a connected network which gets learned based on any machine learning algorithm. The description of the used classifier and the machine learning algorithm in this paper is described in Section 5.1. After the classifier network is made to learn, the test dataset is used to check the performance of the trained classifier network. A k-fold cross validation technique is a statistical approach used to average out the performance of a classifier by repetitively performing the operation on different set of training and test datasets of the same dataset. A *k*-fold cross validation with *k* value equal to 10 is applied and this procedure is made to run for 10 folds. The final performance of the classifier network is calculated as the average of the 10 folds. In this paper, as the focus is on the minority class samples, recall and F-measure are the performance measures which are observed thoroughly. This is

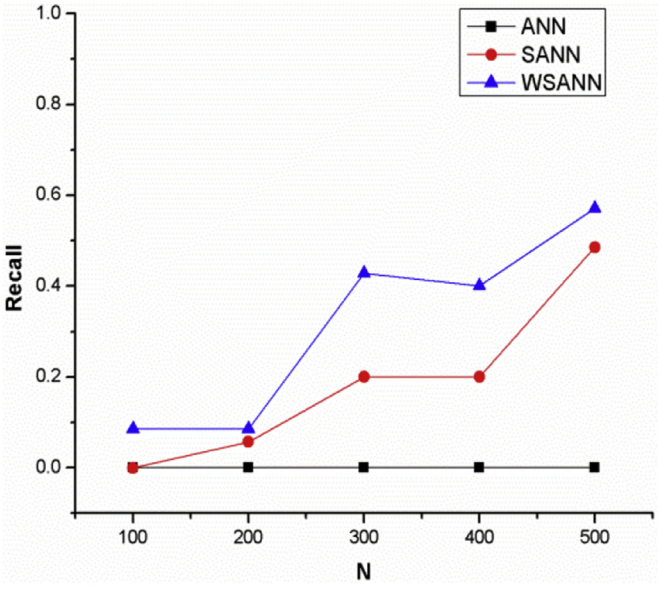

Fig. 4. Comparison of Recall for Ecoli dataset.

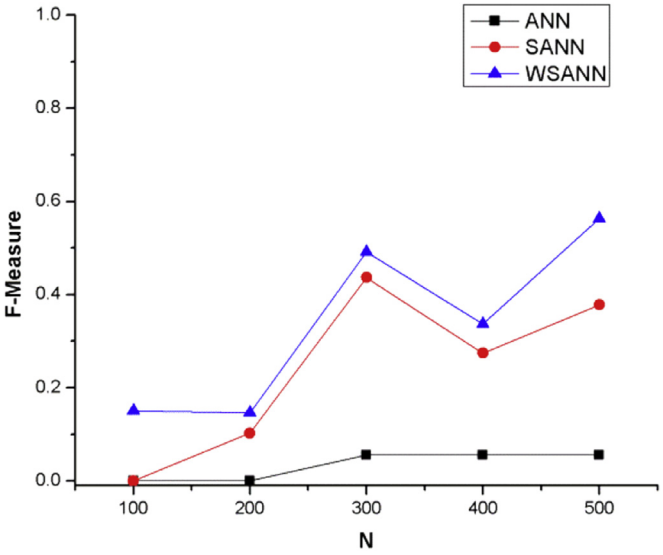

Fig. 5. Comparison of F-Measure for Ecoli dataset.

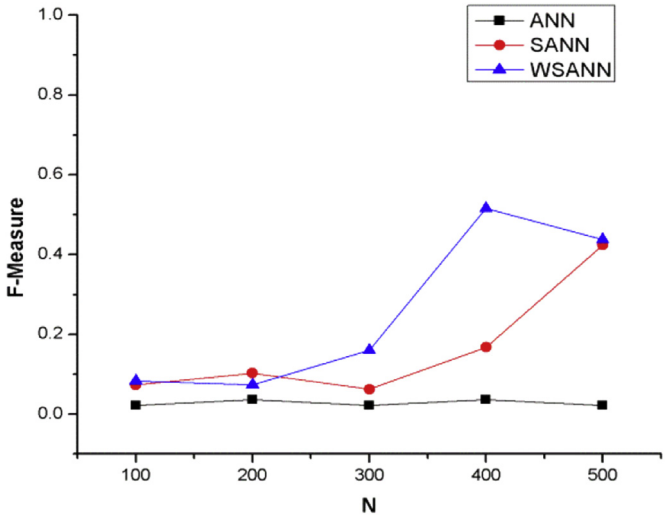

Fig. 7. Comparison of F-Measure for Abalone dataset.

because these two measures explain clearly about the performance of the classifier network towards the minority class.

6. Observation

The five datasets which are considered in this paper as mentioned in Sect. 5 are used to analyze the performance of artificial neural network (ANN) multilayer perceptron in three different ways. The performance measures which are mostly concentrated upon in this paper were recall and f-measure. In the first case, the performance of the classifier is calculated when the dataset is directly fed as input to the ANN classifier without any preprocessing of data denoted as ANN. In the second case, the performance of the classifier network is calculated when the dataset undergoes SMOTE before being fed as input to the ANN classifier, denoted as SANN. In the last case, the performance of the classifier is calculated when the dataset undergoes WSMOTE before being fed as input to the ANN classifier denoted as WSANN. The performance is calculated for different amounts of oversampling (N), i.e., 100%, 200%, 300%, 400% and 500%.

From Figs. 4–13, it is evident that WSANN is placed above ANN in all the figures. Hence the former is certainly better than the later. For the Ecoli dataset, the WSANN gave higher values compared to SANN for recall as well as F-measure in all the five values of N as shown in Figs. 4 and 5. The comparison shown in Fig. 6 shows that WSANN gives higher recall for Abalone dataset compared to all the five different N values using SANN. However, the F-measure for 200% oversampling for WSANN is marginally low compared to SANN shown in Fig. 7. This is because Abalone dataset require a minimum of 300% oversampling to give higher F-measure using WSANN than SANN which is evident from Fig. 7. In case of Wine quality dataset, the recall and F-measure values for WSANN are nearly equal to or slightly higher than that of SANN as shown in Figs. 8 and 9. The recall and F-measure values for the Yeast dataset using WSANN is comparably higher than SANN in all the five cases

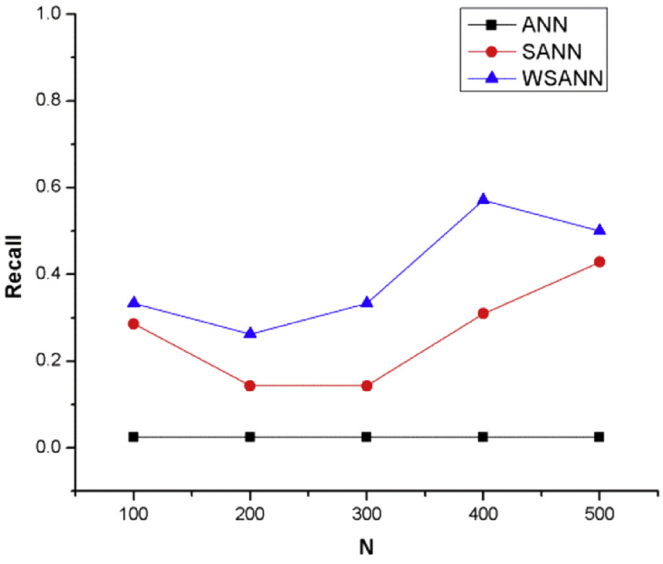

Fig. 6. Comparison of Recall for Abalone dataset.

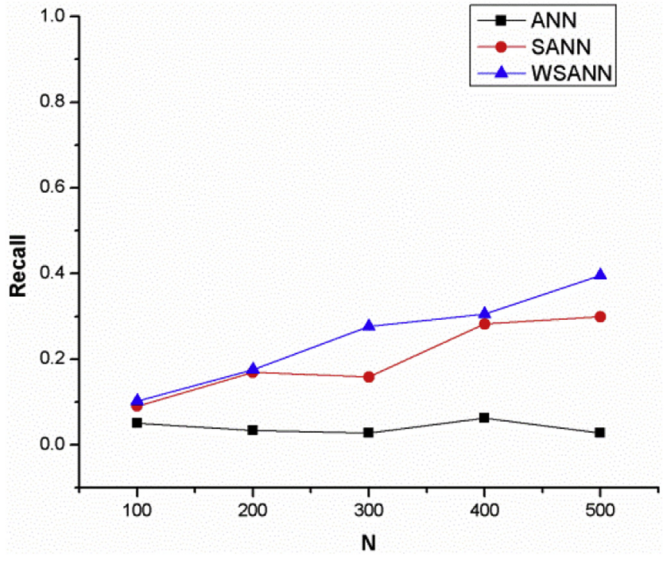

Fig. 8. Comparison of Recall for Wine Quality dataset.

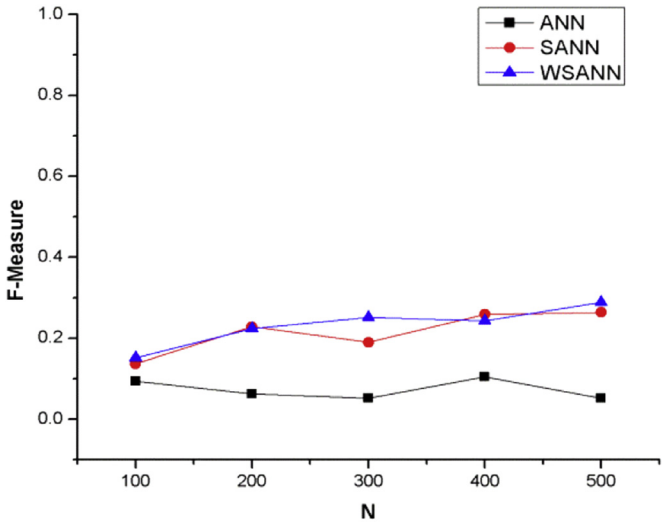

Fig. 9. Comparison of F-Measure for Wine Quality dataset.

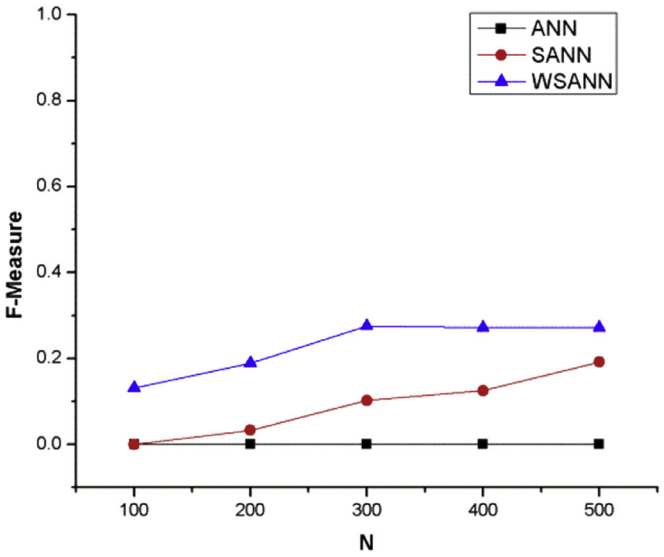

Fig. 11. Comparison of F-Measure for Yeast dataset.

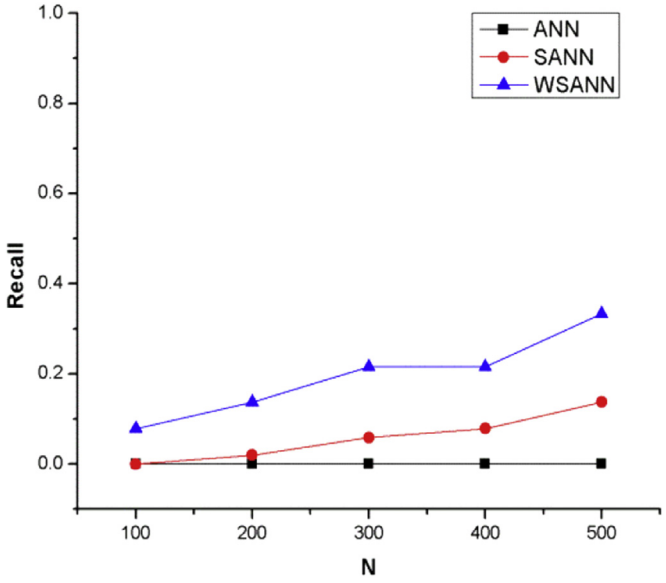

Fig. 10. Comparison of Recall for Yeast dataset.

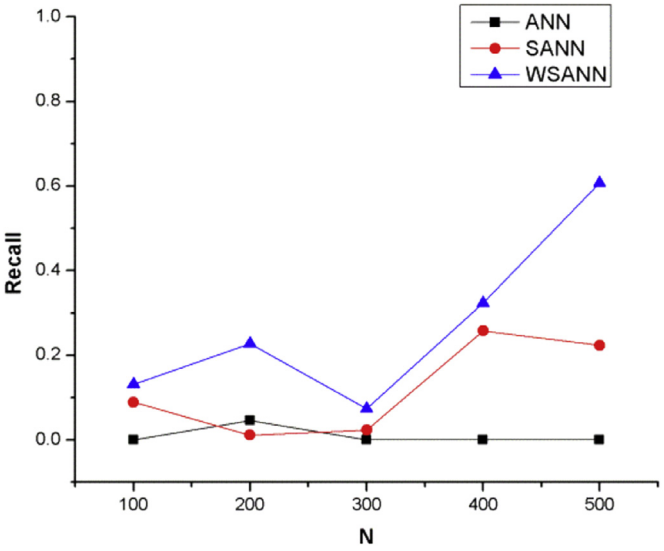

Fig. 12. Comparison of Recall for Mammography dataset.

of oversampling values as shown in Figs. 10 and 11. Figs. 12 and 13 shows that both the recall and F-measure using WSANN is higher than SANN for the Mammography dataset.

7. Comparative analysis on SMOTE and WSMOTE based classifier on imbalanced dataset in a SFR

The performances of both WSMOTE and SMOTE are assessed using some of the malfunction events in a SFR. The data is collected from operator training simulator of Prototype Fast Breeder Reactor (PFBR) which is a SFR being established in Kalpakkam, India. The block diagram of the complete power cycle of PFBR is shown in Fig. 14. Two of the events that are subjected to imbalanced dataset are selected for testing purpose from the training simulator which has been highlighted in Fig. 14. Here also, recall and f-measure are considered as the performance metrics for the three classifiers, i.e., ANN classifier, SMOTE based ANN classifier and WSMOTE based ANN classifier.

The boiler feed pump (BFP) shown in Fig. 14, is present in the steam water side of a SFR. This pump feeds sub-cooled feed water to steam generator at a pressure of 180 bar. This pumping system consists of two 50% turbo-driven BFP and one 50% motor driven BFP which takes over on loss of any of the two turbo-driven BFP. The feed water flow speed is one of the characteristics features of the BFP. The operating feed water flow at 100% is 561 kg/s. The reduction in the speed of the feed water in BFP causes a malfunction which needs to be addressed. Hence, the first malfunctioning event of the SFR which is considered in this paper is the feed water speed reduction to the steam generator up to a certain percentage which occurs due to malfunctioning of the boiler feed pump located in the steam water system. This dataset contains data for the reduction in the feed water speed by 10%, 20%, 30% and so on till 90% where each of these percentage reductions is denoted as a class. Hence, this dataset consisted of 9 classes and a total of 3451 samples. The features which are considered for each sample consisted of seven plant parameters. These are feed water flow, feed

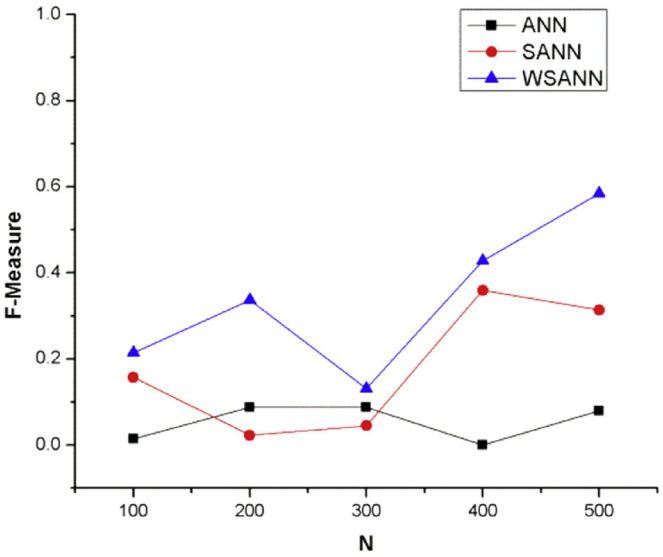

Fig. 13. Comparison of F-Measure for Mammography dataset.

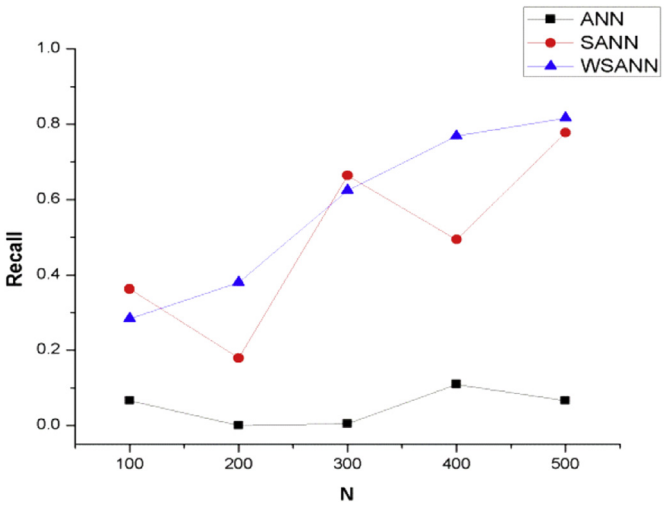

Fig. 15. Comparison of Recall for feed water speed dataset.

water inlet temperature, sodium inlet temperature to steam generator, sodium outlet temperature from steam generator, secondary sodium pump-1 speed, secondary sodium pump-2 speed and superheated steam outlet temperature from steam generator. The reduction in the feed water speed to 90% is considered as the minority class containing 229 samples which has the least count compared to others. Hence, the ratio of minority samples to majority samples is nearly 1:14. Figs. 15 and 16 show that at 200%, 400% and 500% oversampling of minority dataset, WSANN produces higher recall and f-measure respectively compared to SANN. Moreover, both SANN and WSANN produce higher values in both the figures than ANN.

The second malfunctioning event from SFR which had imbalanced dataset is related to secondary sodium pump (SSP) shown in Fig. 14. The SSP pumps in cold secondary sodium from the SG into the IHX. There are two SSP and four IHX in PFBR (one SSP each for two IHX). The speed of SSP at 100% flow is 900 rpm at an operating temperature of 355 °C. The reduction in SSP speed due to its malfunctioning is the second event which is considered in this paper.

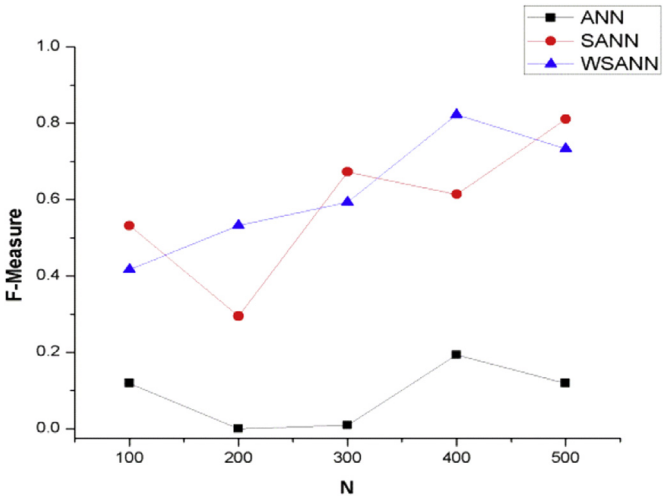

Fig. 16. Comparison of F-Measure for feed water speed dataset.

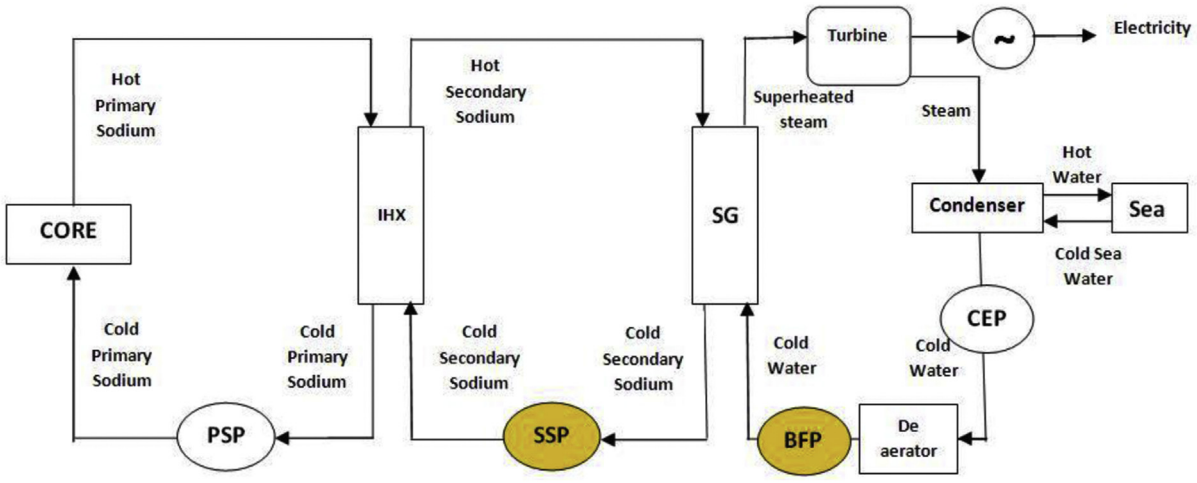

Fig. 14. Block diagram of the flowchart of PFBR.

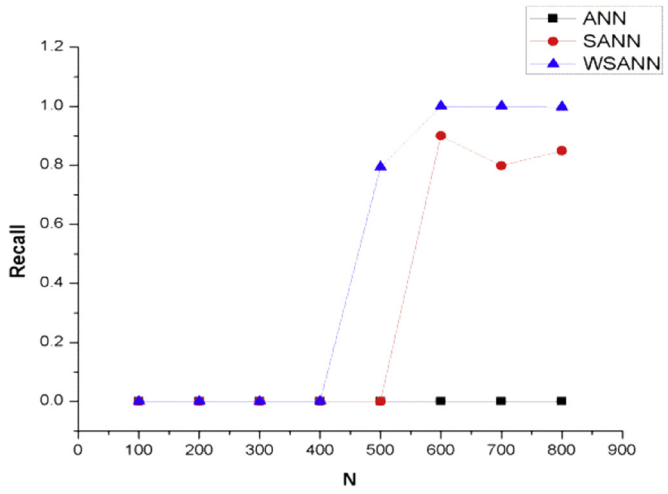

Fig. 17. Comparison of Recall for SSP speed dataset.

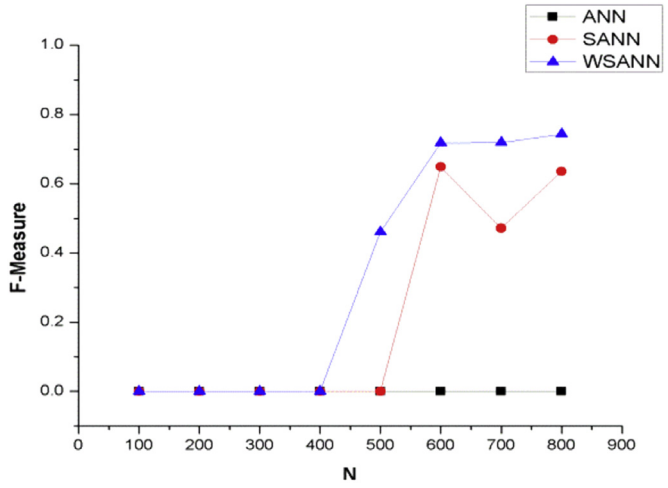

Fig. 18. Comparison of F-Measure for SSP speed dataset.

The dataset considered for this event consisted of data for reduction in SSP speed by 10%, 20%, 30% and 40%. Here, reduction in SSP speed beyond 40% is not considered as the reactor goes to shutdown state. Hence, this dataset consisted of 4 classes and a total of 7785 samples. The same seven plant parameters mentioned in the previous case are considered as features for each sample in this case also. The reduction in SSP speed by 40% is considered the minority class with 708 samples which is again the least count among the others. Hence, the ratio of minority samples to majority samples in this dataset is nearly 1:10. It is observed in Figs. 17 and 18 that both recall and f-measure are zero till the 400% oversampling of minority samples because this amount of oversampling was not sufficient to classify the event. Hence, the experiment is carried out till 800% oversampling for better analysis. It is quite evident from both the figures that both recall and f-measure are higher in case of WSANN compared to SANN. Again, oversampling using SANN and WSANN quite evidently produces better performance than ANN for classifying minority samples.

8. Conclusion

Imbalanced dataset problem and the classifier network learning from this imbalanced dataset are real bottleneck for a machine learning researcher. There is always a need to improve the accuracy of the classifier network for the minority data in such kind of imbalanced datasets. One way of solving this issue is by oversampling the minority data samples. Out of the various oversampling methods, SMOTE had gained a lot of attention. In this paper, a modification to this oversampling technique has been introduced which is termed as Weighted-SMOTE. The Weighted-SMOTE approach which used the Euclidean distance of each minority data sample with respect to the other minority data samples to find the weight associated with each minority data sample produced better performance than SMOTE various test data including that of a SFR simulator. The performance measures which were considered in this paper were recall and F-measure because these two metrics give proper information on the performance of the classifier network towards the minority dataset. The authentication of this approach in diverse fields is yet to be explored. The effect of increase in dimensionality on classifier performance using WSMOTE is certainly the scope of future work.

Acknowledgement

The authors express their sincere thanks to the PFBR Operator Training simulator (KALBR-SIM) team members for providing constant guidance and support in completing this research. The authors are greatly indebted to the constant support and motivation provided by Dr. S. A. V. Satya Murty, Director, Indira Gandhi Centre for Atomic Research. The first author thanks Department of Atomic Energy for the Research Fellowship.

References

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.
Brown, I., Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39, 3446–3453.
Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C., 2009. Safe-level-smote: safe-level-synthetic minority over-sampling TEchnique for handling the class imbalanced problem. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (Eds.), Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 475–482.
Chairi, I., Alaoui, S., Lyhyaoui, A., 2012. Intrusion Detection based Sample Selection for imbalanced data distribution. In: 2012 Second International Conference on Innovative Computing Technology (INTECH), pp. 259–264.
Chawla, N.V., 2005. Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer, US, pp. 853–867.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 321–357.
Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W., 2003. SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (Eds.), Knowledge Discovery in Databases: PKDD 2003, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 107–119.
Cieslak, D.A., Chawla, N.V., Striegel, A., 2006. Combating imbalance in network intrusion datasets. In: 2006 IEEE International Conference on Granular Computing, pp. 732–737.
Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., Bontempi, G., 2014. Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl. 41, 4915–4928.
Estabrooks, A., Jo, T., Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20, 18–36.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett., ROC Anal Pattern Recognit. 27, 861–874.
Fawcett, T., Provost, F., 1997. Adaptive fraud detection. Data Min. Knowl. Discov. 1, 291–316.
Gao, M., Hong, X., Chen, S., Harris, C.J., 2011. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neurocomputing 74, 3456–3466.
Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (Eds.), Advances in Intelligent Computing, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 878–887.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.
He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on

Neural Networks, 2008. IJCNN 2008, pp. 1322–1328.
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284.
Japkowicz, N., 2001. Concept-learning in the presence of between-class and within-class imbalances. In: Stroulia, E., Matwin, S. (Eds.), Advances in Artificial Intelligence, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 67–77.
Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.
Li, J., Li, H., Yu, J.-L., 2011. Application of random-SMOTE on imbalanced data mining. In: 2011 Fourth International Conference on Business Intelligence and Financial Engineering (BIFE), pp. 130–133.
Nahar, J., Imam, T., Tickle, K.S., Shawkat Ali, A.B.M., Chen, Y.-P.P., 2012. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. Expert Syst. Appl. 39, 12371–12377. http://dx.doi.org/10.1016/j.eswa.2012.04.045.
Peng, L., Zhang, H., Yang, B., Chen, Y., 2014. A new approach for imbalanced data classification based on data gravitation. Inf. Sci. 288, 347–373.
Sun, T., Zhang, R., Wang, J., Li, X., Guo, X., 2013. Computer-aided Diagnosis for Early-stage Lung Cancer Based on Longitudinal and Balanced Data.
Yu, H., Ni, J., Zhao, J., 2013. ACOSampling: an ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. Neurocomputing 101, 309–318.
Zeng, Z.-Q., Gao, J., 2009. Improving SVM classification with imbalance data set. In: Leung, C.S., Lee, M., Chan, J.H. (Eds.), Neural Information Processing, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 389–398.
Zhai, Y., Ma, N., Ruan, D., An, B., 2011. An effective over-sampling method for imbalanced data sets classification. Chin. J. Electron 20, 489–494.