International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

An Oversampling Mechanism for Multimajority Datasets using SMOTE and Darwinian Particle Swarm Optimisation

Rose Mary Mathew[1], Dr. R. Gunasundari[2]
[1]Research Scholar, Department of Computer Science,
Karpagam Academy of Higher Education **,**
Coimbatore, India
rosem.mathew@gmail.com
[2]Professor, Department of Computer Applications,
Karpagam Academy of Higher Education**,**
Coimbatore, India
gunasoundar04@gmail.com

Abstract— Data skewness continues to be one of the leading factors which adversely impacts the machine learning algorithms performance. An approach to reduce this negative effect of the data variance is to pre-process the former dataset with data level resampling strategies. Resampling strategies have been seen in two forms, oversampling and undersampling. An oversampling strategy is proposed in this article for tackling multiclass imbalanced datasets. This proposed approach optimises the state-of-the-art oversampling technique SMOTE with the Darwinian Particle Swarm Optimization technique. This proposed method DOSMOTE generates synthetic optimised samples for balancing the datasets. This strategy will be more effective on multimajority datasets. An experimental study is performed on peculiar multimajority datasets to measure the effectiveness of the proposed approach. As a result, the proposed method produces promising results when compared to the conventional oversampling strategies.

Keywords- imbalanced, optimization, oversampling, multiclass, multimajority

I. INTRODUCTION

For the past few years, skewed data has been considered as the sole arduous issue in the domain of data science and machine learning. This skewed issue occurs in cases where there is a great difference in the number of instances of the classes reviewed. This skewness results in poor generalization and causes a learning bias towards the majority classes [1]. Data imbalance is wide ranging in many applications including medical diagnosis, biological data analysis, financing, fraud detection, neuroimaging and so on [2]. Owing to its ubiquitous, this skewed data encountered a prevalent consideration from the research community. Imbalanced data issues in the field of binary classification, multiclass classification, big data, and data streams have been considered in recent years and still there are some issues that remain unsolved.

To resolve problems of data skewness in binary classification, there are two approaches that should be performed. Data level and algorithm level solutions are the different approaches to resolve this issue. Data level approach is carried out via methods of undersampling, oversampling or combination of both. These methods balance the distribution of the data among the majority classes and minority classes. Algorithm level approach is carried out via means of changing the classifier techniques or optimizing the overall conduct of the learning algorithms [3]. The benefits of resampling can be used independent of the selected classifier.

Multi-class imbalanced concerns are observed as considerably more troublesome than the binary partners for various reasons. Skewness can appear in various ways in the case of a multi-class dataset. Multi-class skewness can be either one minority class with a couple of majority classes (Multimajority cases) or one majority class with a couple of minority classes (Multiminority cases)[3]. It is difficult to make an accurate prediction from the multiclass imbalanced datasets. Many techniques are available to tackle this issue. For handling multiclass imbalanced data, class decomposition techniques are used. In this technique multiple classes are decomposed into combinations of binary classes and handle the imbalanced issue. This technique of partitioning has different approaches. Two popular approaches are One against All and One against One. These two techniques are combined to form a hybrid approach termed All and One. Data level approaches can be applied on the existing dataset to minimize the effect of majority classes

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

and minority classes [4].By using the resampling techniques, the issue of skewness can be minimized.

In this article, a sophisticated approach is put forward to deal with multimajority datasets. The proposed approach performs a resampling over the minority class of the multimajority dataset. Resampling is done in the form of oversampling the minority class data with the aid of an optimization technique [5]. The strategy proposed in this article is formulated on the belief of generating optimized synthetic samples for the minority classes to balance the dataset. The major contributions of this article can be summarized in this way; proposition of a sophisticated optimized oversampling technique termed DOSMOTE and exploratory assessment of the proposed method.

The rest of this article is systematized as follows. In the second section, discussion of the significant scientific contributions in handling data skewness was specified. In the third section the proposed approach is introduced in detail. Experimental study including comparison of results are presented in the fourth section. Ultimately, the article is wrapped up with major findings and future work in the fifth section.

II. Related Works

In this section, the different procedures are discussed for handling the imbalanced datasets. Imbalanced data can be handled by different methods either resampling the dataset or using algorithmic solutions or by cost sensitive measures. Most of the studies are concentrated on the resampling techniques. Resampling techniques change the order of data distribution, and the changes are independent of the elemental classifier. Resampling approaches are assorted into three groups; they are oversampling, undersampling and hybrid methods [5]. In oversampling method, a superset of the original dataset will be created by producing synthetic data points for the minority class. In case of undersampling, a subspace of the initial dataset will be created by wiping out the instances of the majority classes. In hybrid approach, it combines the basics of oversampling and undersampling approaches to get the dataset balanced.

A lot of survey papers are available in skewed data learning. Sun et al. give an outline about the imbalanced data problems and discuss the various strategies to resolve the same[6]. A good intuition into the essence of skewed data distributions are discussed by Lopez et al. [7]. Galar et al. give a survey report of implementation of ensembled classifiers to the skewed data [8]. Wang and Yao have focused on multi-class imbalance data learning [9]. Zhang et al., in their work, decomposes the multi-class imbalance data with the aid of one-versus-one method and balances it for model creation [10]. Problems of imbalanced data is well discussed by Krawczyk[11]. Skewed data consistently do not proffer a problem aside however connected with alternative data controversial factors; it skeptically affects the acceptance of the minority class and the same is effectively discussed by Stefanowski [12].

Oversampling paradigms are mainly focused on neighbourhood-based policies and the same is originated from the Synthetic Minority Over-sampling Technique (SMOTE) which is recommended by Chawla [13]. Later the original SMOTE algorithm is blended with some boosting techniques and developed SMOTEBoost algorithm[14]. Another advancement on SMOTE is safe-level-SMOTE proposed by Bunkhumpornpat it considers the opposite approach and it focuses on safest instances[15]. Maciejewski and Stefanowski have proposed LN-SMOTE which manipulates the details about the neighbourhoods of oversampled instances. MWMOTE, an expansion of SMOTE by growing the synthetic sample generation procedure based on clustering mechanism was proposed by Sukarna Barua. An approach for determining the safety of an instance based on its local neighbourhood and the same was proposed by Napierała [16].

An important extension of SMOTE algorithm is the ADASYN which was proposed by H.He et al. [17]. SPIDER algorithm has been proposed by Stefanowski and Wilk recognizes the local aspects of the samples, and then eliminates those majority datapoints which causes misclassification of samples from the minority data[18]. The neighbourhood cleaning rule by Laurikkala eliminates difficult instances which depends on their vicinity[19]. EasyEnsemble and BalanceCascade by Liu et al. combine undersampling with ensembled classifiers [14]. Undersampling in sync with the evolutionary algorithms are proposed by Garcia and Herrera [20] and the same is later expanded in the form of EUSBoost. Fernandez-Navarro et al. recommend a widening of SMOTE algorithm to the multi-class scenario [21].

Some of the hybrid resampling techniques, that is the possibility of bringing together oversampling and undersampling is investigated by Estabrooks. Batista et al. propose to use SMOTE by combining some methods for data cleaning called Tomek links [22]and the Edited Nearest-Neighbor rule(ENN) [23]. Later several methods are proposed by combining oversampling with undersampling of the given samples.

III. Proposed Work

A lot of techniques were proposed by researchers for imbalanced data handling and these resampling techniques achieved satisfactory results over skewed datasets. Oversampling methods have settled to be further competent than other techniques, however it often destroys the distribution of actual data. SMOTE produces fabricated samples of the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

minority data by calculating the distance between minority sample and its closest neighbor, however the sample space of the majority class is often invaded by the newly generated synthetic samples. The newly created synthetic sample will also influence the following data process, which will affect the results of classification process. To get better of this issue, in the proposed work applying optimization technique to the newly generated synthetic samples. In this proposed work Synthetic Minority Oversampling Technique (SMOTE) and Darwinian Particle Swarm Optimization (DPSO) are combined to a single algorithm for getting the optimized balanced data.

The proposed work focuses on oversampling of the data. To make the dataset balanced, oversampling methods keep the original data of all the majority classes and add synthetic data to the minority classes. In this work the multimajority datasets are focused. Multimajority datasets mean for a multiclass dataset, the data distribution of multiple classes is almost same and high. Only a few classes have low frequency of data. Multiple high frequency availed classes are termed as majority classes and low frequency data availed classes are termed as minority classes. This work is focused on generating synthetic data of minority class and to minimize the alteration of original dataset. This method produces remarkable results on multimajority datasets.

To solve the skewed data problems, the commonly used oversampling method is SMOTE. Its goal is to balance the distribution of data by putting synthetic data points towards the minority class. It blends new minority class samples between surviving minority samples. It produces synthetic instances through linear interpolation of the minority data. These fabricated instances are produced by randomly choosing one or more of the k numbers of the closest neighbors for every sample in the minority data [13]. For generating synthetic samples using SMOTE, set up minority class as X, for every data x in X, the k numbers of its closest neighbors of x are identified by considering the Euclidean distance between x and every other data in set X. According to the proportion of skewness, the sampling rate N is to be fixed. For each x in X, N samples are randomly selected from the k number of its nearest neighbors and construct the set. For each sample x_i in the collection of closest neighbors in minority class, the following formula is used to produce new data,

$$y = x + rand(0,1) * (x - x_i)$$

where rand (0,1) represents a random number between 0 and 1[13]. The newly generated synthetic data is added to the dataset and is used for further classification processes. However, this newly generated synthetic sample sometimes overlaps with the existing majority class data, and this will affect the classification process. So, to resolve this issue an optimization technique is introduced to optimize the newly generated data. The optimization technique using in this work is Darwinian Particle Swarm Optimization (DPSO) [24].

Optimization techniques are used in Machine Learning to obtain better results. Particle Swarm Optimization (PSO) is one of the probabilistic optimization techniques that is focused on the flow and skill of swarms(bits). It utilizes several bits that add up to a swarm which is moving in the search space and gazing for the best results. Every bit in PSO has their own background of reaching the optimum solution. It connects with neighbourhood bits to assure even if its values for appropriate results are optimum or not. It found that if its neighboring bit has better standards, then it seeks to get close to those standards. For the whole swarm, a global best value is saved and each bit in the swarm seeks to attain this global best value. This is an iterative mechanism and is repeated except for some trivial advances identified in the global best value. PSO faces an issue of stagnation as bit pace and positions are randomly chosen and updated which can point to standards that will not generate optimum results [25].

In multiclass imbalanced data classification, multiple classes are present in the dataset. If synthetic samples are generated using SMOTE, then it will be checked with the existing data of multiple majority classes. And the same can be performed by the optimization technique. In Particle Swarm Optimization, it can work with a single swarm at a time. So, this work focused on Darwinian Particle Swarm Optimization (DPSO). In the DPSO model, collective swarms can be taken in sync to the same problem and act as a sole PSO algorithm. The bit whose efficacy is not able to obtain optimum solution, or it induces the complete swarm to stagnation and the same is tested at each iteration and if noticed not fit for farther execution then it is eliminated from the swarm and life of the swarm is reduced. The whole swarm will be eliminated when swarm life attains its minimum [24].

In Darwinian PSO, collective swarms with test results may live at every pace. Every swarm singly acts like a simple PSO algorithm which has some protocols guiding the group of swarms that sketched to mimic the natural selection. This natural selection process implemented as a choice of swarms in a continually changing group of swarms. In this work by using SMOTE, artificial samples of minority classes are produced based on the difference among the existing minority data and its closest neighbor. However, the sample space of the majority data class is often invaded by the newly generated samples of minority data. This newly generated sample will also affect the subsequent data process, which will influence the classification result. To overcome this issue, optimization of the newly generated samples is to be done[26]. By the application of DPSO to the newly generated samples it becomes optimized. In

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

this proposed work the combination of SMOTE and DPSO is used for getting the balanced data.

The pseudocode of the proposed method takes the original imbalanced dataset as the input. It first identifies the count of classes listed in the dataset. And the count of instances presents in each class. Based on the data, it identifies the minority and majority classes. Then it invokes the DOSMOTE method. In this method it initializes the particles and parameters in DPSO with initial values. For generating synthetic samples, it identifies the nearest neighbors of an existing instance of the minority class. Artificial samples will be created based on the equation given below.

$$X_{syn} = x_i + rand\ (0,1) * |x_i - x_{near}|$$

Where x_i *-> Minority sample*
x_{near} *-> Nearest neighbor of* x_i *in the minority data*

The generated samples are moved to the DPSO structure, which uses the evolutionary process to optimize the particle (synthetic sample generated) globally, which is varied from the neighbourhood information that is the data present in the other classes. The other classes are treated as different swarms. For each swarm, based on the newly generated sample fitness values are identified. Based on that, it finds the best data and updates the position and velocity of the particle. If any particle or sample is obsolete, it discards the same. This process will be repeated until it reaches the iteration value. Only the best particles will be moved to the final balanced dataset. After doing all the iterations optimized synthetic samples are obtained which makes the dataset balanced as the output. The Pseudo code of the proposed work DOSMOTE is present in Table 1.

Table 1. Algorithm for DOSMOTE

Pseudo code for DOSMOTE

Input: Imbalanced dataset
Output: Balanced Dataset

1. Identify the count of different classes present in the given dataset.
2. Identify the count of instances present in each class.
3. Detect the minority and the majority classes.
4. Perform DOSMOTE Oversampling technique.
   1. Initialize the particles and parameters in DPSO
   2. For every instance present in the minority class do
      Identify the nearest neighbors of each minority instance.
      Generate sample by using following equation.
      $X_{syn} = x_i + rand\ (0,1) * |x_i - x_{near}|$
      Where x_i *-> Minority sample*
      x_{near} *-> Nearest neighbor of* x_i *in the minority data*
      End
   3.Initialize the global best and velocity of the swarm
   4. For each swarm in the collection, find the fitness values based on the newly generated data
   5. Find the global best data.
   6. Update position value and velocity of the particle.
   7. Repeat steps 4 to 6 until the loop reaches the iteration value.
   8. Select the generated minority class sample based on global best particle.

IV. Experimental Study

This study is related to analyzing the efficiency of various resampling techniques collectively with the advanced machine learning algorithms over multiclass skewed data. The input for the study is extracted from the Keel data repository and UCI Machine Learning repository. Implementation of this work is done in MATLAB. The different stages of this work are represented in the form of a figure. Figure.1 shows the workflow representation of this experiment. The different steps that are followed in this study are,

i. Selection of dataset from the KEEL Repository. Three multimajority datasets of 3-class classification are chosen.
ii. Selected datasets have undergone data pre-processing stage to remove the noise data.
iii. Split the dataset in the ratio 70:30 for performing training and testing.
iv. Evaluate the performance of classifiers like k-NN, SVM and ANFIS.
v. Apply resampling techniques DOSMOTE, SMOTE and ADASYN to the training dataset.
vi. After applying each resampling technique, it evaluates the performance of the various classifiers like k-NN, SVM and ANFIS.
vii. Analyze the performance of these classifiers before applying resampling technique and after applying the resampling technique.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

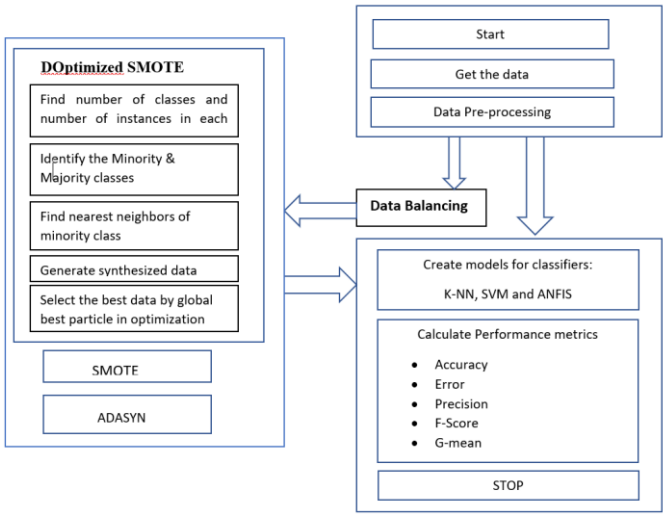

Figure 1. Describing the various stages in the experimental study.

A. Data used for study.

A list of experiments is assisted to study the effectiveness of the prospective algorithm DOSMOTE. The empirical review is conducted on multiclass skewed datasets available in the KEEL and UCI data repositories[27][28]. For this study, multimajority datasets are focused. Multimajority dataset means multiple majority classes together with a few minority classes. Three multimajority datasets Balance, Hayes-Roth and Web-Phishing dataset are chosen for the study. The information related to these datasets is given in table, Table 2. The original data distribution plots of the three datasets are shown in the figure, Figure 2.

1. Resampling methods used for comparison.

Resampling methods are enforced in the data pre-processing phase. For this study, the proposed work DOSMOTE applied datasets, classification results are compared with the existing oversampling techniques like Random Oversampling, SMOTE and ADASYN methods.

2. Random Oversampling (ROS)

ROS is the lightest form of oversampling technique which is used in the data pre-processing stage. In this method the data from the minority classes are selected randomly and replicated to make a balance with the count of the majority class data [4]. This is applicable for all minority classes in the multiclass classification problem. Since the existing points are replicated there will be a possibility for increasing the overfitting of data.

3. SMOTE

SMOTE is an oversampling technique which produces synthetic datapoints alternatively photocopy the live datapoints. To generate new datapoints k-nearest neighbor method is used. The value of k in k-nn depends on the number of new datapoints created for making the dataset balanced. The distance between feature vector and neighboring points are calculated with any of the available distance formulas [13]. The variation in the distance is noted for different points and this variation is multiplied with a random value in the set (0,1). The value of the product is included in the feature vector as the new data value.

Table 2. Dataset Information

Name of the dataset	No. of Instances	Imbalance Ratio	No. of Attributes	Class Labels	No. of records
Balance	625	5.88	5	Balance	49
				Left	288
				Right	288
Hayes-Roth	160	2.1	5	1	51
				2	51
				3	30
Web-Phishing	1353	6	11	-1	702
				0	103
				1	548

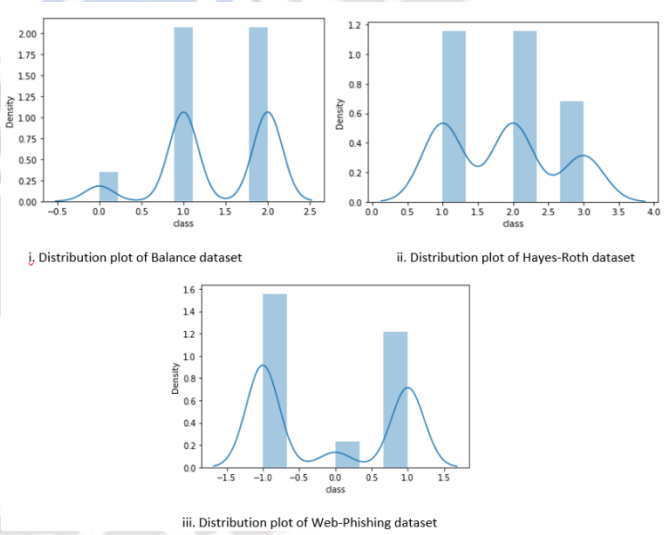

Figure 2. Data distribution of various datasets

4. DOSMOTE

The pseudocode specified in Algorithm.1 works. The parameters used for the proposed work are specified in Table 3.

Table 3. Parameters used in the Optimization Method

Particle Number	100
Iteration Number	1000
weight	0.5
Constants (c1, c2)	(1.8, 2)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

B. Classifiers used for study.

For this study, three different classifiers are used to create the models. The different classifiers are K-NN (K-Nearest Neighbors), SVM (Support Vector Machines) and ANFIS.

1. K-NN

K-Nearest Neighbor follows the lazy learning technique. This classifier performs well for predictive analysis[29]. In this technique for the test data 'k' number of closest neighbors are identified and the classes are identified for these k neighbors and the class occurs with high frequency is fixed as the class of test data.

2. Support Vector Machines

SVM is the most popularly utilized classifier for doing the classification. In SVM the algorithm generates a decision boundary for the data points and this decision boundary is called hyperplane. This hyperplane is made in such a manner that it keeps maximum distance away from the data values. This hyperplane is termed as maximum margin hyperplane (MMH).

3. ANFIS

ANFIS follows neuro-fuzzy strategy. It is a mix of neural networks and fuzzy logic which has been announced to tame the flaws and to provide better engaging features. The intention of this classifier is to eliminate non-specific information available in data and provides results which is described by high interpretability and have gained an extent of accuracy. This system sketches the input aspects to different input membership functions and after that it sketches input membership functions to different rules and then these rules to a collection of output aspects. Finally, it sketches output aspects to output membership functions, and the output membership function to a sole output or a judgment linked with the output. In this system a neural network architecture is being used to depict the input/output in case it sketches inputs through input membership functions and related criterions, and after that the output membership functions and related criterions to outputs, The criterions related with membership functions revisits the previous learning process [30].

C. Performance Measures.

The different models are developed with a few classifiers. The exhibition of these models is to be assessed. For doing the assessment, confusion matrices are utilized. Confusion matrices are based on actual values and predicted values. If the prediction and actual data are same, it is termed as correct classification otherwise it is termed as incorrect classification. The values in confusion matrix are True-Positive (TP), False-Positive (FP), True-Negative (TN), False-Negative (FN). By using these values, accuracy of the model can be predicted. This experiment is on imbalanced datasets so that the performance can be evaluated with some other metrics like Accuracy, Error, F1-Score, Precision, and G-mean.

Accuracy is termed as the correctness of the model created. The values in confusion matrix are used for calculating accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Error of the model is obtained by taking the difference of percentage of accuracy from the whole.

$$Error = 1 - Accuracy$$

Precision of a model is defined as the rate at which samples are correctly classified. It is termed as the true positive rate.

$$Precision = \frac{TP}{TP + FP}$$

F1-Score is a measure which is obtained by combining precision and recall values. This is evaluated as the harmonic mean between recall and precision [29].

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

G-mean is obtained by taking the square root of the product of sensitivity and specificity. For multi-class data the same is obtained by the higher root of the product of sensitivity for each class.

$$G - Mean = SQRT(Sensitivity * Specificity)$$

D. Results and Discussions

For this experimental study, a multimajority of datasets are used. All the selected datasets have three class labels. So, the predicted results will be one from any of these three classes. In the multiclass classification process the evaluation is done by the confusion matrix. The confusion matrix generated by the datasets are of the form 3X3. Performance metrics values are considered for all the class combinations. For the easiness of assessment, the average produced by the results is considered.

The performance of the model can be evaluated by practicing the confusion matrix. For each dataset the order of the confusion matrix is 3*3, as the dataset has three class entries. The evaluation metrics like precision, F-score, G-mean and accuracy will be evaluated for these multiple classes. The detailed evaluation is presented in the following tables. The following Table 4. shows the accuracy, error, precision, F-Score and G-Mean of the various classes by applying the different classifiers on the original imbalanced dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
Article Received: 16 December 2022 Revised: 14 January 2023 Accepted: 30 January 2023

Table 4. Performance metrics of different imbalanced datasets

Name of the dataset	Name of the classifier	Accuracy	Error	Precision	F-Score	G-Mean
	K-NN	0.8085	0.1915	0.5625	0.5834	0.5837
	SVM	**0.8776**	**0.1224**	**0.585**	**0.6194**	**0.6205**
Balance	ANFIS	0.7991	0.2009	0.7097	0.6948	0.8895
	K-NN	**0.5128**	**0.4872**	**0.6815**	**0.5194**	**0.5963**
	SVM	0.4615	0.5385	0.5392	0.4956	0.5907
Hayes-Roth	ANFIS	0.4615	0.5385	0.5166	0.4671	0.5587
	K-NN	0.8128	0.1872	0.7511	**0.7019**	**0.7148**
	SVM	**0.8152**	**0.1848**	**0.7661**	0.6316	0.6916
Web-Phishing	ANFIS	0.5852	0.4148	0.6043	0.5252	0.6743

Table 5. Distribution of data over various classes before and after resampling

	Imbalanced data distribution			After Oversampling								
				ROS/SMOTE			ADASYN			DOSMOTE		
Dataset	**class1**	**class2**	**class3**	**class1**	**class2**	**class3**	**class1**	**class2**	**class3**	**class1**	**class2**	**class3**
Balance	49	288	288	233	233	233	240	233	231	117	186	186
Hayes-Roth	51	51	30	41	41	41	41	41	42	35	35	35
Web-Phishing	702	103	548	565	565	565	565	565	435	567	480	566

In the data preprocessing stage various oversampling methods mentioned above are used for balancing the dataset. The experiment uses a data split of ratio 70:30 towards training and testing of the model. The distribution of data used for training before and after application of resampling techniques are represented in Table 5. In the case of Random Oversampling and SMOTE, both produce the same number of samples after resampling. The resampling techniques make the datasets almost balanced. In the case of oversampling, it performs oversampling in minority classes. All the selected datasets have multiple majority classes, so the resampling should be happened on the minority class.

To assess the effectiveness of DOSMOTE technique, check the performance metrics of different datasets with different resampling techniques and classifiers. Firstly, considering the assessment of Balance dataset. Table 6. shows the performance metric values for Balance dataset in case of different resampling situations over various classifiers. From Table.5 the performance metrics of Balance dataset, the classification model with oversampling technique DOSMOTE, together with the classifier KNN achieved the better performance. Accuracy, Precision, F1-Score and G-mean are high for DOSMOTE resampled dataset with K-NN classifier model. For ANFIS classifier, better results are produced with the DOSMOTE technique. Figure.3 shows the performance chart of Balance dataset over various classifiers.

Hayes-Roth is another multimajority dataset that is in account. Table 7. shows the performance metric values for Hayes-Roth dataset in case of different resampling situations over various classifiers. For the Hayes-Roth dataset the classification model with resampling technique ROS, SMOTE and ADASYN together with SVM algorithm produces the better performance. Support Vector Machine algorithm produces same result for all the resampled datasets expect DOSMOTE resampling technique. However, Other classifiers like K-NN and ANFIS models produce better results with DOSMOTE resampling technique. Figure.4 shows the performance chart of Hayes-Roth dataset over various classifiers.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

Table 6. Performance metrics of Balance dataset.

Name of the resampling technique	Name of the classifier	Accuracy	Error	Precision	F-Score	G-Mean
Imbalanced Set	K-NN	0.8085	0.1915	0.5625	0.5834	0.5837
	SVM	0.8776	0.1224	0.585	0.6194	0.6205
	ANFIS	0.7991	0.2009	0.7097	0.6948	0.8895
ROS	K-NN	0.734	0.266	0.588	0.5801	0.5812
	SVM	0.8989	0.1011	0.8374	0.8524	0.8739
	ANFIS	0.71	29	0.57	0.549	0.5444
SMOTE	K-NN	0.755	0.245	0.687	0.674	0.6804
	SVM	0.8723	0.1277	0.8167	0.8229	0.8532
	ANFIS	0.67	0.33	0.698	0.625	0.6873
ADASYN	K-NN	0.75	0.25	0.6691	0.6567	0.6628
	SVM	0.8829	0.1171	0.8242	0.8338	0.8607
	ANFIS	0.656	0.344	0.616	0.57	0.59
DOSMOTE	K-NN	**0.9023**	**0.0977**	**0.9061**	**0.9162**	**0.9417**
	SVM	0.8762	0.1238	0.792	0.8276	0.9267
	ANFIS	0.8567	0.1433	0.7626	0.7773	0.9208

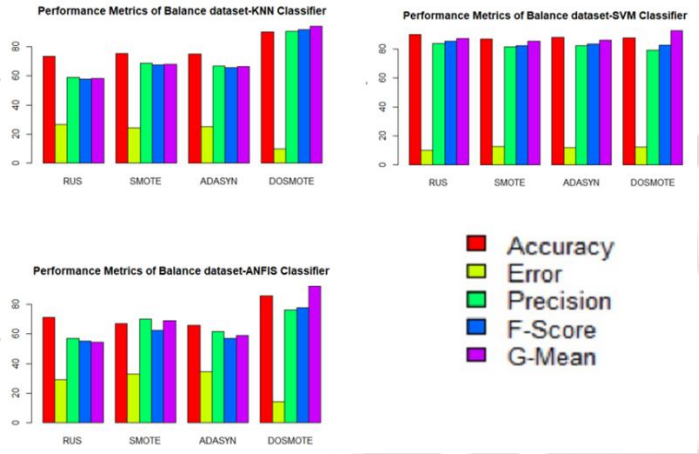

Figure 3. Graph showing the performance of Balance dataset.

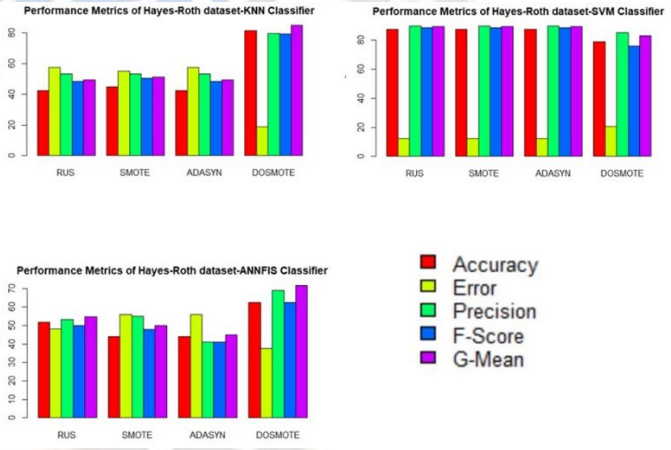

Figure 4. Graph showing the performance of Hayes-Roth dataset.

Table 7. Performance metrics of Hayes-Roth dataset.

Name of the resampling technique	Name of the classifier	Accuracy	Error	Precision	F-Score	G-Mean
Imbalanced Set	K-NN	0.5128	0.4872	0.6815	0.5194	0.5963
	SVM	0.4615	0.5385	0.5392	0.4956	0.5907
	ANFIS	0.4615	0.5385	0.5166	0.4671	0.5587
ROS	K-NN	0.425	0.575	0.5343	0.4855	0.4933
	SVM	**0.875**	**0.125**	**0.8981**	**0.8877**	**0.8934**
	ANFIS	0.518	0.482	0.533	0.5	0.5492

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

	K-NN	0.45	0.55	0.533	0.5049	0.5104
	SVM	**0.875**	**0.125**	**0.8981**	**0.8877**	**0.8934**
SMOTE	ANFIS	0.44	0.56	0.55	0.479	0.5013
	K-NN	0.425	0.575	0.5343	0.4855	0.4933
	SVM	**0.875**	**0.125**	**0.8981**	**0.8877**	**0.8934**
ADASYN	ANFIS	0.44	0.56	0.41	0.41	0.4527
	K-NN	0.8125	0.1875	0.7953	0.7928	0.8495
	SVM	0.7917	0.2083	0.852	0.7614	0.8312
DOSMOTE	ANFIS	0.625	0.375	0.6916	0.6249	0.7174

Table 8. Performance metric values for Web-Phishing dataset

Name of the resampling technique	**Name of the classifier**	**Accuracy**	**Error**	**Precision**	**F-Score**	**G-Mean**
	K-NN	0.8128	0.1872	0.7511	0.7019	0.7148
	SVM	0.8152	0.1848	0.7661	0.6316	0.6916
Imbalanced Set	ANFIS	0.5852	0.4148	0.6043	0.5252	0.6743
	K-NN	0.7832	0.2168	0.7033	0.7232	0.7391
	SVM	0.8275	0.1725	0.7581	0.7823	0.7935
ROS	ANFIS	0.69	0.31	0.63	0.62	0.6249
	K-NN	0.7857	0.2143	0.7053	0.7241	0.7446
	SVM	0.83	0.17	0.7629	0.7833	0.7891
SMOTE	ANFIS	0.66	0.34	0.597	0.579	0.5879
	K-NN	0.7586	0.2414	0.6792	0.688	0.7164
	SVM	0.8226	0.1774	0.7505	0.7722	0.7835
ADASYN	ANFIS	0.58	0.42	0.525	0.515	0.5455
	K-NN	**0.9409**	**0.0591**	**0.9083**	**0.9238**	**0.9551**
	SVM	0.9051	0.0949	0.8425	0.8658	0.9276
DOSMOTE	ANFIS	0.4895	0.5105	0.6527	0.4786	0.6746

The third multi-majority dataset under consideration is Web-Phishing dataset. Table 8. shows the performance metric values for Web-Phishing dataset in the case of different resampling situations over various classifiers. For the Web-Phishing dataset the classification model K-NN with the proposed resampling technique DOSMOTE produces the better result. The SVM classifier also has got the best results with the DOSMOTE resampled datasets. However, the performance of ANFIS is not up to expectations. Figure 5. shows the performance chart of Web-Phishing dataset over various classifiers.

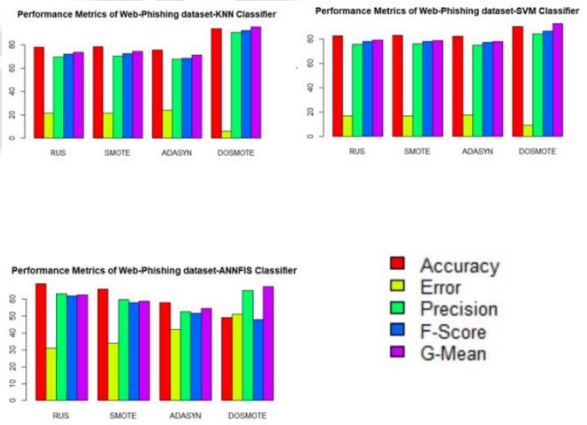

Figure 5. Graph showing the performance of Web-Phishing dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

The performance metrics of various resampled dataset with different algorithms are already specified. From the metric tables the proposed oversampling technique DOSMOTE technique produces promising results for all the datasets. Even in Hayes Roth dataset proposed resampling technique performs better with KNN and ANFIS classifiers than other resampled datasets. Oversampling techniques preserve all the data that are available in the dataset and in the case of proposed method it optimizes the generated synthetic samples, and these optimized samples are considered for model creation.

Based on the experimental results presented it is evident that DOSMOTE technique outperforms the existing oversampling algorithms. The proposed method obtained the highest values when compared with most of the considered methods. The proposed optimized oversampling technique has achieved utmost performance respective to the referred oversampling strategies on different datasets. This indicates that the proposed strategy is pertinent for dealing with imbalanced datasets. The experimental review also refers that in some exceptional cases like Hayes-Roth dataset existing oversampling technique together with a particular classifier SVM has obtained better performance however other classifier's performance is remarkable about the proposed algorithm.

V. Conclusion

In this paper, a sophisticated approach is proposed for tackling data skewness by preserving the original data distribution. The proposed approach oversamples the minority classes of skewed dataset. For preserving the distribution of original dataset, this algorithm is proposed to multimajority datasets, so that only a few minority classes will be affected by synthetic samples. In this proposed approach all the artificial samples are optimized by the Darwinian Particle Swarm Optimization technique so that these artificial samples will not overlap with the existing data, so that model performance is increased. Experimental results presented in this article indicate that DOSMOTE method surpasses the subsisting resampling methods, in sync with several classifiers. To tackle multi-class imbalanced data over multimajority datasets, the proposed algorithm will produce promising results. A direction for further research is to develop an adequate algorithm to tackle multi-class skewed data over multiminority datasets. Another research direction is used to extend this optimized oversampling method to the big data scenario to tackle the multiple class skewed classification problems.

References

[1] W. Liu et al., "A comprehensive active learning method for multiclass imbalanced data streams with concept drift," *Knowledge-Based Syst.*, vol. 215, p. 106778, 2021.

[2] M. Koziarski and M. Wozniak, "CCR: A combined cleaning and resampling algorithm for imbalanced data classification," *Int. J. Appl. Math. Comput. Sci.*, vol. 27, no. 4, pp. 727–736, 2017, doi: 10.1515/amcs-2017-0050.

[3] S. S. Yadav and G. P. Bhole, "Learning from Imbalanced Data in Classification," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 1907–1916, 2020.

[4] R. M. Mathew and R.Gunasundari, "A review on handling multiclass imbalanced data classification in education domain," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE*, 2021, pp. 752–755, doi: 10.1109/ICACITE51222.2021.9404626.

[5] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 42, no. 4, pp. 1119–1130, 2012, doi: 10.1109/TSMCB.2012.2187280.

[6] Sun et al. (2009). Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence* **23**(04): 687–719.

[7] L´opez et al. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, *Information Sciences* **250**: 113–141.

[8] M. Galar et al., "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognit.*, vol. 46, no. 12, pp. 3460–3471, 2013.

[9] Wang, S. and Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions, *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics* **42**(4): 1119–1130

[10] Zhang et al. (2016). Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data, *Knowledge-Based Systems* **106**: 251–263.

[11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[12] Stefanowski, J. (2016). Dealing with data difficulty factors while learning from imbalanced data, *in* S. Matwin and J. Mielniczuk (Eds.), *Challenges in Computational Statistics and Data Mining*, Springer, Heilderberg, pp. 333–363.

[13] Chawla et al. (2002). SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* **16**: 321–357.

[14] Chawla et al. (2003). SMOTEBoost: Improving prediction of the minority class in boosting, *European Conference on Principles of Data Mining and Knowledge Discovery, Cavtat/Dubrovnik, Croatia*, pp. 107–119.

[15] Bunkhumpornpat et al. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, *Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand*, pp. 475–482.

[16] Napierała, K. and Stefanowski, J. (2012). Identification of different types of minority class examples in imbalanced data, *International Conference on Hybrid Artificial Intelligence Systems, Salamanca, Spain*, pp. 139–150.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6139
***Article Received**: 16 December 2022 **Revised**: 14 January 2023 **Accepted**: 30 January 2023*

[17] He, H et al. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE International Joint Conference on Neural Networks (IEEEWorld Congress on Computational Intelligence), Hong Kong, China*, pp. 1322–1328.

[18] Stefanowski, J. and Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance, *International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy*, pp. 283–292.

[19] Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution, *Conference on Artificial Intelligence in Medicine in Europe, Cascais, Portugal*, pp. 63–66.

[20] Garcia, S. and Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy, *Evolutionary Computation* **17**(3): 275–306.

[21] F. Fernández-Navarro et al., "A dynamic over-sampling procedure based on sensitivity for multi-class problems," *Pattern Recognit.*, vol. 44, no. 8, pp. 1821–1833, 2011, doi: 10.1016/j.patcog.2011.02.019.

[22] Tomek, I. (1976). Two modifications of CNN, *IEEE Transactions on Systems, Man, and Cybernetics* **6**(11): 769–772.

[23] Wilson, D.L. (1972). Asymptotic properties of nearest neighbor rules using edited data, *IEEE Transactions on Systems, Man, and Cybernetics* **2**(3): 408–421.

[24] J. Tillett et al., "Darwinian particle swarm optimization," *Proc. 2nd Indian Int. Conf. Artif. Intell. IICAI 2005*, pp. 1474–1487, 2005.

[25] Kennedy J. and Eberhart R. C. Particle swarm optimization. In *Proceedings of the International Conference on Neural Networks*; Institute of Electrical and Electronics Engineers. Vol. 4. 1995. pp. 1942–1948. DOI: 10.1109/ICNN.1995.488968

[26] Xi-Bin Dong Xian-Bing Meng Zhi-Wen Yu Philip Chen Guo-Qiang Han, "A PSO-optimized Oversampling Method for Imbalance Classification", Key-Area Research and Development Program of Guangdong Province No. 2018B010107002,2020

[27] J. Alcalá-Fdez *et al.*, "KEEL: A software tool to assess evolutionary algorithms for data mining problems," *Soft Comput.*, vol. 13, no. 3, pp. 307–318, 2009, doi: 10.1007/s00500-008-0323-y.

[28] J. Alcalá-Fdez *et al.*, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2011.

[29] Rose Mary Mathew, Dr. R. Gunasundari. (2021). An Experimental Study on The Effect of Resampling Techniques in Multiclass Imbalanced Data in Learning Sector. *Design Engineering*, 16216-16234. Available at http://www.thedesignengineering.com/index.php/DE/article/view/6768

[30] Hosseini, M. S., & Zekri, M. (2012). Review of Medical Image Classification using the Adaptive Neuro-Fuzzy Inference System. *Journal of medical signals and sensors*, *2*(1), 49–60.

Author Biography

Mrs. Rose Mary Mathew is a Research Scholar in the Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore. Currently she is working as Assistant Professor (Special Grade) in the Department of Computer Applications, Federal Institute of Science and Technology, Angamaly. She has more than ten years of teaching experience. She obtained her Master of Computer Applications degree from Mahatma Gandhi University, Kottayam in 2009 and MBA from Bharathiyar University, Coimbatore in 2018. Her area of specialization is Machine Learning. She had attended several National and International seminars and conferences.

Dr.R. Gunasundari is presently working as Professor and Head of the Department, in the Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore. She has more than fifteen years of teaching experience. She has participated and presented several papers in National and International conferences. Her research interests include Data Mining, Cryptography and Network Security.