Journal of Computer Science

Original Research Paper

Data Analytics for Imbalanced Dataset

Madhura Prabha R and Sasikala S

Department of Computer Science, University of Madras, Chennai, India

Article history
Received: 27-01-2023
Revised: 04-07-2023
Accepted: 19-07-2023

Corresponding Author:
Madhura Prabha R
Department of Computer Science, University of Madras, Chennai, India
Email: madhura.prabha@gmail.com

Abstract: The primary issue in real-time big data classification is imbalanced datasets. Even though we have many balancing techniques to reduce imbalance ratio which is not suitable for big data that has scalability issues. This study is envisioned to explore different balancing techniques with experimental study. We tried comparing the effectiveness of various balancing strategies, including cutting-edge approaches for severely unbalanced data from online repositories. Here we apply SMOTE, SMOTE ENN and SMOTE Tomek balancing algorithms for dermatology, wine quality and diabetes datasets. After balancing the dataset, the balanced dataset is classified with AdaBoost and random forest algorithms. On three datasets, the outcomes show that the classification algorithm with the balancing technique improves the classification performance for imbalanced datasets. Experiment results showed that the SMOTE ENN technique produces higher classification with accuracy than the SMOTE and SMOTE Tomek techniques. The findings are analyzed with other factors like execution time and scalability. Though SMOTE Tomek produces 1.0 for a few datasets, its execution time is longer than SMOTE ENN. Therefore, SMOTE ENN with random forest classification produces 1.0 accuracy for all three datasets with less execution time. This experimental study analyses to create a novel ensemble technique for balancing highly imbalanced data.

Keywords: Imbalanced Dataset, Balancing, Irregular Case, Scalability, Multi-Class, SMOTE, SMOTE Tomek, SMOTE ENN

Introduction

An imbalanced dataset increases the bias in classification, especially for finding irregular cases like disease identification, leakage finding, machine-fault identification, fraud detection, etc., With the continuous capturing of data from many big data sources like video surveillance, satellite images, social media data and finance transactions, it is very difficult to infer knowledge about data (He and Garcia, 2009). Rare events like cancer gene detection, natural disasters, machine faults, oil spill detection and fraudulent credit card transactions are hard to find because of their rarity and informality (Haixiang *et al*., 2017). The sparse occurrences of rare events will make the dataset imbalanced. i.e., In a skewed dataset, the interested class or the abnormal class occurrences are very few than the normal class occurrences.

Identifying minority instances from majority instances and precisely obtaining essential information is difficult in skewed data. New imbalanced data problems have emerged as a result of the development of big data and machine learning (Haixiang *et al*., 2017).

Traditional classifiers' results may go wrong due to a concentration on classes that are high in count and ignoring the minority instances (Zhao *et al*., 2021). Generally, classifiers predict low accuracy with an imbalanced dataset (Batista *et al*., 2004). In classification or regression methods, the result shows bias towards the normal class or majority class which has a huge number of occurrences and that ignores rare class or minority class. Bias problems can be solved by applying a suitable balancing algorithm for a classifier to balance a dataset from an imbalanced one.

Normally an imbalanced dataset contains numerous instances of non-interested class and very few instances of interested class. This will produce imbalanced training data as well as a poor classifier model. Hence classification becomes a challenging task in an imbalanced dataset (Fernández *et al*., 2017).

The class imbalance will cause more errors and bias toward the majority of classes in the dataset. It also produces more false negatives which costs more than false positives. If we use these imbalanced datasets for classification training, it will create a wrong model with only majority classes and neglect minority classes. This will create performance deterioration (Fernández *et al*., 2017).

Minority class identification is a big issue in classification tasks. Training models can recognize the majority class only and produce the wrong output. This problem will make the data more complex (Bader-El-Den *et al*., 2018).

There are different imbalanced attributes that will affect the classification results. Lin and Chen (2013) identify the imbalanced qualities as (i) The uniqueness of both classes, (ii) The ratio of the minority class size to that of the majority class and (iii) The shortage of training data. The ratio of imbalance is reflected in the first property. If the skewness increases the classification error increases. The second attribute is class size. These problems will be solved by balancing minority and majority instances. The last attribute depends on the first two attributes. When the count of the minority class is less, it lacks minority information. Identifying the boundary between two different classes is difficult which leads to low performance in minority class prediction.

Data level balancing, algorithm level balancing and ensemble methods are some of the approaches available for balancing imbalanced data. The data level balancing approach will modify dataset instances with certain strategies. The algorithm method will adjust the existing classifier to get higher classification accuracy. The modification considers misclassification costs which are called cost-sensitive methods which will minimize cost error instead of increasing accuracy (Tanha *et al*., 2020). The ensemble technique will incorporate both data and algorithm-level techniques to attain high accuracy in classification.

In imbalanced datasets, when the counts of the majority and minority classes diverge significantly, most balancing techniques target binary classes. Multi-class unbalanced data focuses on the dominant classes rather than the minority classes (Sleeman IV and Krawczyk, 2021). Binarization splits multi-class into many binary classes which will lose a lot of important information.

Big data brings with it new classification-related issues and difficulties. The challenges of volume, velocity, variety, veracity and value are met by big data (Fernández *et al*., 2017). The next challenge is scalability, which can be addressed by developing new techniques and solutions for big data scenarios. Spark has arisen as a widespread method to develop models on big data (Zaharia *et al*., 2012). Map reduce programming style is used to adapt big data classification. It follows standard techniques and partitioning which will create small disjuncts and loss of data (Fernández *et al*., 2017).

This study aims to experiment with different balancing techniques for three different datasets which are highly imbalanced. After balancing, classification can be done by AdaBoost classification and random forest classification algorithms. Finally, results are compared based on different criteria.

This study has various sections. The literature review section describes the related work done in imbalanced data and various balancing techniques. The methodology section defines the detailed view of different balancing techniques. The experimental work section defines the datasets used and experimental results. The performance results are tabulated. The conclusion section summarizes the entire work.

Nowadays, the popular method to predict and analyze business values from the existing dataset is the classification method. Each domain has a different variety of datasets. A skewed dataset is an imbalanced one that has two different classes. The following are two different types:

i. Majority class: Class which has a greater count of instances
ii. Minority class: Class which has a lesser count of instances

The imbalanced dataset can be categorized into two categories:

i. An imbalanced dataset with two classes: A majority and a minority
ii. An imbalanced dataset with multiple classes: "n" majority and "n" minority classes (Fernández *et al*., 2018)

The sampling method can be rearranged to get good predictions (Amrehn *et al*., 2018). The classifier performance is very low in imbalanced training data.

The problems in imbalanced classification are small disjuncts, noisy data and borderline issues (Ramyachitra and Manikandan, 2014). Imbalanced classification is a more critical issue than binary class imbalanced learning. There are different ensemble techniques that increase machine learning performance (Tanha *et al*., 2020). The imbalance problem can be categorized as within labels, between labels and among the label sets (Tarekegn *et al*., 2021).

Binary Class Balancing Techniques

The following are three different techniques to balance the imbalanced dataset:

1) Data level technique: It adds or removes a few data instances depending on the problem domain. Tanha *et al.* (2020) distinguished between two types of sampling techniques: Under and oversampling
2) Algorithm level approach: It modifies the classifier algorithm to balance the dataset
3) Ensemble Techniques: It combines both data and Algorithm approaches. Ensembles use a divide-and-conquer method to increase classification performance. Ensemble methods combine different simple techniques (Hasanin and Khoshgoftaar, 2018)

Combining both techniques is called hybrid sampling. Hybrid sampling selects samples that are more significant and removes samples that are least significant (Susan and Kumar, 2021).

The synthetic minority oversampling technique (SMOTE), which replicates the original datasets using synthetic instances, was introduced by Chawla *et al*. (2002). SMOTE produces a preference for lower-count classes and emphasizes them. Many minority instances are created near existing samples. Rather than creating data space points, feature space data points are created. The synthetic samples are created by oversampling minority classes. The KNN method is used in oversampling.

Batista *et al*. (2004) proposed SMOTE and Tomek link which removes majority class samples that make Tomek links and minority class samples are also removed. Imbalanced datasets are balanced using SMOTE + Tomek or SMOTE + ENN (Noorhalim and Shamsuddin, 2019) with fewer positive instances which leads to classification performance accuracy. When positive samples are high in number, the random oversampling method is suitable and less expensive.

A Balanced Random Forest (BRF) with combined sampling and ensemble techniques was proposed by Bader-El-Den *et al*. (2018). Samples are synthetically created and rearranged to equally split into different individual trees. Weighted random forest assigns more weight to the minority class and less weight to the majority class and a penalty for misclassifying class. Sampling and ensemble learning methods are combined to change the class size and balance all classes. Experiment results show that both weighted RF and balanced RF are better than existing techniques.

Ning *et al*. (2021) suggested the DEXGB_Glu method, which uses the XGBoost classifier based on the differential evolution algorithm to identify lysine glutarylation sites. It is a hybrid of Tomek and borderline-SMOTE. The differential evolution technique improved performance and solved the balancing problem between majority samples and minority samples. The performance was better prediction methods when compared with other methods of glutarylation sites.

Fernando and Tsokos (2021) proposed a class rebalancing technique that dynamically assigns weights based on class frequency. Experiments conducted on intrusion detection and medical imaging datasets. Results based on theory with the help of superior empirical performance give verification of dynamically weighted balanced loss function. The Dynamically Weighted Balanced (DWB) function is supported by experimental results.

Juez-Gil *et al*. (2021) experimented with ensemble techniques conducted for imbalanced datasets, using bagging and boosting techniques. Experiments were done in a spark environment. Results and execution time were compared with the Bayesian approach. The conclusion is an interesting one in that simpler methods give better results than complex methods for imbalanced datasets. Because of its complexity, it is not a good technique to balance imbalance for normal-sized datasets. Pre-processing technique is an essential step and it should be done at the beginning or end of a training session of ensemble technique.

Zhu *et al*. (2018) proposed class weights random forest to balance the imbalanced dataset. This technique can identify both classes with good accuracy which shows that it increased the entire performance of classification.

Li *et al*. (2019) proposed a unified data-preparation method using stochastic swarm heuristics to increase and optimize both majority and minority classes by reproducing the training dataset. This method produces better results than other methods.

Genetic algorithms served as the foundation for a proposed classifier for an unbalanced dataset (GAs). Principal Component Analysis (PCA) examined datasets and found errors. By their method, the mistakes in a dataset were displayed in binary form. Through GA, error location identification was accomplished. The imbalanced dataset was processed more quickly than the GA-based technique, which had been successful in pinpointing the error's source.

The Neighbor Cleaning Rule (NCL), as introduced by (Hasanin and Khoshgoftaar, 2018), entails finding each example's k = 3 closest neighbors to improve the ENN technique for two-class issues. A majority class instance will be eliminated if it has a prediction fault along with one of its closest neighbors (Hasanin and Khoshgoftaar, 2018). If a neighbor is a member of the minority class and there is a prediction error involving them, the closest neighbors who are in the majority class will be removed.

By carefully balancing the data for the performance of diagnostics in the medical field, the issue can be resolved without much difficulty. Junsomboon and Phienthrakul (2017) proposed an imbalance dataset adjustment method through the integration of the synthetic minority over-sampling technique (SMOTE) and Neighbor Cleaning rule (NCL) methodologies (Popel *et al*., 2018).

Viola and Jones (2001) use the AdaBoost algorithm to pick a few critical visual attributes from potential attributes. It provides strong ties to generalization results and an efficient learning algorithm. The goal of the AdaBoost algorithm is to identify a limited number of highly diverse, high-quality features. The minority class was able to set the feature using this method. Single features, or single-node decision trees, are used by the weak classifiers (An and Kim, 2010).

Multi-Class Balancing Techniques

Multi-class balancing can be performed by ensemble binarization techniques. Binarization is performed by decomposition methods (Lorena *et al*., 2008). "One-Vs-One" (OVO) (Knerr *et al*., 1990) and "One-Vs-All" (OVA) (Clark and Boswell, 1991) are common decomposition techniques.

An issue of class 'm' is divided into binary problems of class (m*(m-1))/2 using the OVO decomposition method. All of them are resolved using binary classifiers (Fernández *et al*., 2018). Only a portion of the original training dataset's instances with one of the two corresponding class labels are used to train the classifier; instances with different class labels are ignored (Fernández *et al*., 2018).

An m-class problem is further divided into m-binary problems by the OVA method. A binary classifier is used in each problem to distinguish between the two classes (Fernández *et al*., 2018). Using all of the training data, the classifier is trained to treat every class pattern as positive and every other example as negative.

The bagging method considers classifier variations, performance, classifier count and training sets but in imbalanced classification balancing should be done before classification (Roshan and Asadi, 2020).

The common framework for parallel processing is map reduction. Map reduce is open-source and overcomes scalability issues. It effectively uses a "divide-and-conquer" strategy to be fault-tolerant and adjust to common hardware (Fernández *et al*., 2017).

An unbalanced stream can change the role of labels. A majority class can become a minority class and vice versa. These phenomena make the development of resampling-based stream ensembles difficult. Observed changes in the measurement behavior can be directly transferred to the efficiency of the drift detector that monitors the unbalanced current. Measurements are measured by a normalization method to increase awareness of possible changes in values caused solely by evolving class relations (Brzezinski *et al*., 2019).

Materials and Methods

This section describes the imbalanced classification challenges and their prevalence results. Creating classifiers based on ensemble approaches is an effective way to solve this problem (Zhao *et al*., 2021). Imbalanced datasets are collected from online repositories, pre-processing is done on the dataset. Then the classifier takes a dataset with balanced classes as input and classifies the given balanced dataset.
The methodology has four modules as follows:

1. Big data collection
2. Big data pre-processing
3. Balanced big data classification
4. Evaluation

The proposed model is shown in Fig. 1. Datasets are collected from the Kaggle data repository. The datasets for diabetes, dermatology and wine quality were selected due to their unbalanced Ratios (IR) between majority and minority classes (Fernández *et al*., 2018).

Big Data Collection

i. Dermatology dataset: In dermatology, differential diagnosis of erythematous squamous cell disease is a serious issue. Erythema and scaling are clinical features in the dermatology data set (Çetin and Gökhan, 2018). In this domain, the constructed dataset has a family history attribute that has a family history attribute that has a family history attribute that has value of if these Diseases are found in the family and 0 or 0. The patient's age is indicated by the age attribute (Çetin and Gökhan, 2018). Clinical and histopathological feature scores range from 0-3. If the feature was not available, the value was 0, the maximum was 3 and the relative median values were 1 and 2
ii. Wine quality dataset: This dataset can be used for classification or regression. The classes are imbalanced and ordered from excellent to poor. Outlier detection algorithms are used to find excellent or poor wines
iii. Diabetes dataset: The national institute of diabetes, digestive and kidney diseases provided this dataset. The goal is to determine if a patient has diabetes by using diagnostic tools

Big Data Pre-Processing

In classification, pre-processing can be extremely important (Kotsiantis *et al*., 2006). It includes various data cleaning processes like data balancing, feature extraction, handling missing values and discretization, etc. This experiment uses data balancing techniques for unbalanced datasets. Under-sampling and over-sampling are the two main categories into which various balancing techniques fall.

Using the under-sampling technique, an unbalanced dataset is cleaned up of a few majority class instances. Thereby it balances the count of positive and negative classes. However, this may have a chance of deleting the important instance from the dataset. This can be avoided by the oversampling technique.

The oversampling technique duplicates a few minority instances in an imbalanced dataset. Thereby it balances the count of positive and negative classes. The following are different oversampling techniques:

1. Random oversampling
2. SMOTE
3. SMOTE Tomek
4. SMOTE ENN

Fig. 1: Proposed model

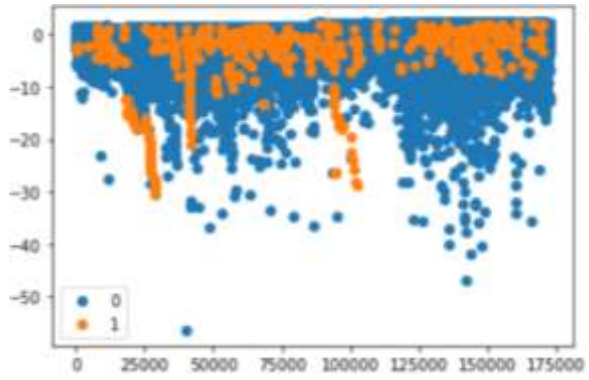

Fig. 2: Feature space before SMOTE

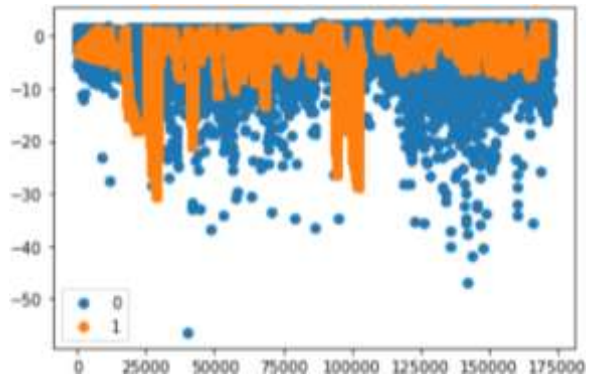

Fig. 3: Feature space after SMOTE

Random oversampling technique duplicates minority instances in an imbalanced dataset. But for duplicating instances, it takes random minority instances. This leads to overfitting in feature space and increases outliers. The SMOTE method is applied to the dataset in order to prevent overfitting and outliers.

SMOTE: Synthetic Minority Oversampling Technique

SMOTE creates synthetic minority instances near existing instances and a larger area is covered by minority classes. This makes classifiers better predict hidden instances of minority classes. SMOTE is an oversampling technique that creates a wide region for minority instances. In this way, SMOTE will play a significant role in the feature space.

Figures 2-3 show feature space before and after the SMOTE technique. SMOTE introduces synthetic minority instances and increases feature space region. However, it introduces additional noise and increases class overlapping. Class overlapping introduces borderline issues. SMOTE Tomek removes borderline issues.

SMOTE Tomek

A SMOTE technique with a cleaning extension is called SMOTE Tomek (Batista *et al*., 2004). Synthetic samples are created and borderline issues are reduced by the Tomek link. Tomek creates synthetic samples and removes borderline instances from feature space. So, it combines both under-sampling and over-sampling techniques. But it still introduces noises in feature space. SMOTE ENN is applied to remove noises.

SMOTE ENN

After SMOTE in the pipeline, SMOTE-ENN is an under-sampling technique that focuses on eliminating noisy samples to produce cleaner combined samples (Li *et al*., 2019). ENN is a data sanitization method to remove samples from both classes. Therefore, samples that are misclassified by neighbors are removed from training data (Batista *et al*., 2004).

Nearest neighbors are computed based on the Euclidean distance of each combined sample. Samples that differ from neighboring samples are removed from the original data set. SMOTE ENN creates a wide feature space with synthetic instances. It also removes borderline issues and noises.

Balanced Classification

Balanced big data may be classified using the AdaBoost classifier and random forest classifier. Classifier performances may depend upon dataset characteristics (Yijing *et al*., 2016).

AdaBoost Classification

AdaBoost classifier is one of the most popular and very strong algorithms (An and Kim, 2010). It is an ensemble classifier that has member classifiers.

The efficiency of this algorithm depends upon the diversity of associate classifiers and their performance. Associate classifiers are selected in the training process to reduce faults in each iteration step.

Random Forest Classification

Multiple decision trees form a Random Forest (RF). The balanced data is given to RF which classifies a dataset with higher performance than an imbalanced dataset.

Matching can be done using one of the SMOTE methods and the resulting matched dataset is passed to the RF classifier. The classifier classifies in an ideal way. Random forest classifiers can achieve high data classification accuracy compared to many standard classifiers, the error rate is minimized and data with imbalanced classes. We have some problems, but the main problem in finance, health care and other fields is class imbalance (Makki *et al*., 2019).

Balancing can be done using any one of the SMOTE methods and the resultant balanced dataset is given to the RF classifier. The classifier will classify in an ideal way. In this method, balancing and classification are two phases. Therefore, the execution time is less than the balanced random forest classifier.

The variety of member classifiers and the algorithm's performance are what determine its performance (An, 2010). During the training process, member classifiers are chosen with the goal of lowering errors at each iteration step.

Results

The proposed methodology is implemented with three datasets. Table 1 presents the attributes of the dataset.

The number of majority class/number of minority class is the imbalanced ratio.

Random forest classification and AdaBoost classification are done on these selected datasets. Metrics such as precision, recall, accuracy and others are used to gauge the outcomes. In terms of precision, the SMOTE ENN approach yields better results. There are metrics like recall, F-score, accuracy and confusion matrix.

The confusion matrix is a useful tool for comparisons. For the test set, projected classes are represented as columns and actual classes as rows. Both balanced and unbalanced data sets are subjected to random forest and AdaBoost classification.

Dermatology Dataset

In AdaBoost classification, SMOTE ENN has 1.0 accuracy. Whereas in random forest classification, SMOTE Tomek and SMOTE ENN both have 1.0 accuracy.

Through experiments, values for F-measure, accuracy, precision and recall are obtained. Figures 4-5 show that SMOTE ENN is performing well when compared with all other methods for the dermatology dataset. The computational results are listed in Table 2.

Wine Quality Dataset

In AdaBoost classification, all balancing methods produce 0.7 accuracy. Whereas in random forest classification, SMOTE TOMEK has got 0.99 accuracy and SMOTE ENN has 1.0 accuracy. This shows that for the wine quality dataset, SMOTE ENN with random forest classification produces better results than all other methods for the wine quality dataset.

Through experiments, values for F-measure, accuracy, precision and recall are obtained. Figures 6-7 show that SMOTE ENN is performing well when compared with all other methods for the wine quality dataset. The computational results are listed in Table 3.

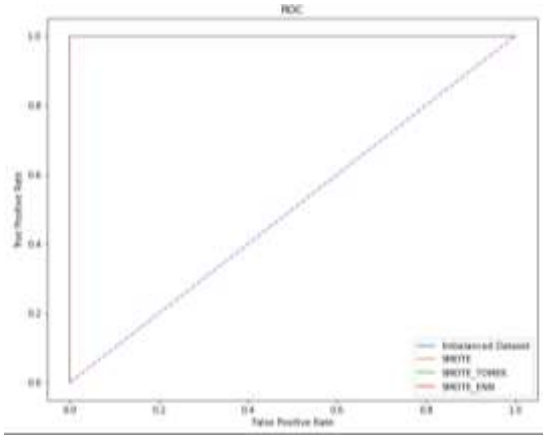

Fig. 4: ROC for AdaBoost classification dermatology dataset

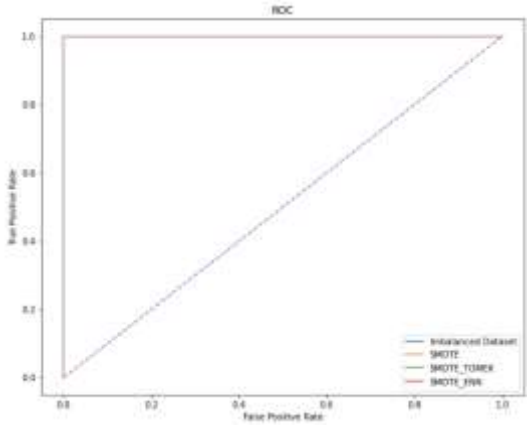

Fig. 5: ROC for random forest classification dermatology dataset

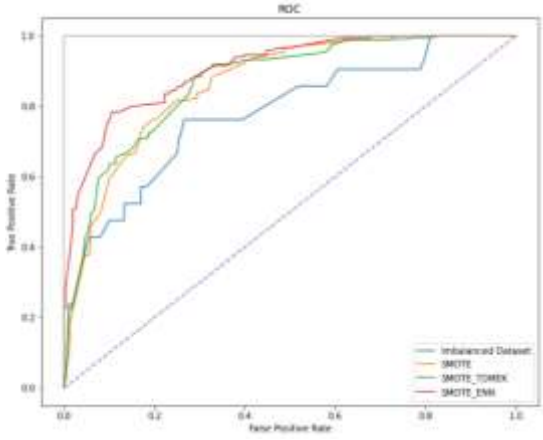

Fig. 6: ROC for AdaBoost classification wine quality dataset

Table 1: Datasets

Dataset	Total instances	No. of attributes	Imbalanced ratio	No. of instances in minority	No. of instances in a majority
Dermatology	58	35	1:17	20	338
Wine quality	1599	12	1:133	53	1546
Diabetes	768	9	1:2	268	500

Table 2: Performance results of the dermatology dataset

Classifier	Balancing technique		Precision	Recall	F-measure	Accuracy
AdaBoost classifier	Original (imbalanced)	0	0.9900	1.00	0.99	0.9950
		1	1.0000	0.99	1.00	
	SMOTE (balanced)	0	0.9900	1.00	0.99	0.9950
		1	1.0000	0.99	1.00	
	SMOTE Tomek (balanced)	0	0.9900	1.00	0.99	0.9950
		1	1.0000	0.99	1.00	
	SMOTE ENN (balanced)	0	1.0000	1.00	1.00	1.0000
		1	1.0000	1.00	1.00	
Random classifier	Original (imbalanced)	0	1.0000	1.00	1.00	0.9994
		1	0.8900	0.70	0.78	
	SMOTE (balanced)	0	1.0000	1.00	1.00	0.9998
		1	0.9998	1.00	1.00	
	SMOTE Tomek (balanced)	0	1.0000	1.00	1.00	1.0000
		1	0.8900	0.70	0.78	
	SMOTE ENN (balanced)	0	1.0000	1.00	1.00	1.0000
		1	0.9998	1.00	1.00	

Table 3: Performance results of wine quality dataset

Classifier	Balancing technique		Precision	Recall	F-measure	Accuracy
AdaBoost classifier	Original (imbalanced)	0	0.96	1.00	0.98	0.9583
		1	1.00	0.05	0.09	
	SMOTE (balanced)	0	0.80	0.74	0.77	0.7704
		1	0.74	0.81	0.77	
	SMOTE Tomek (balanced)	0	0.77	0.76	0.77	0.7647
		1	0.76	0.77	0.76	
	SMOTE ENN (balanced)	0	0.73	0.76	0.75	0.7873
		1	0.83	0.81	0.82	
Random classifier	Original (imbalanced)	0	0.96	0.99	0.98	0.9541
		1	0.40	0.10	0.15	
	SMOTE (balanced)	0	0.94	0.96	0.95	0.9482
		1	0.96	0.94	0.95	
	SMOTE Tomek (balanced)	0	0.94	1.00	1.00	0.9978
		1	0.96	1.00	1.00	
	SMOTE ENN (balanced)	0	1.00	1.00	1.00	1.0000
		1	1.00	1.00	1.00	

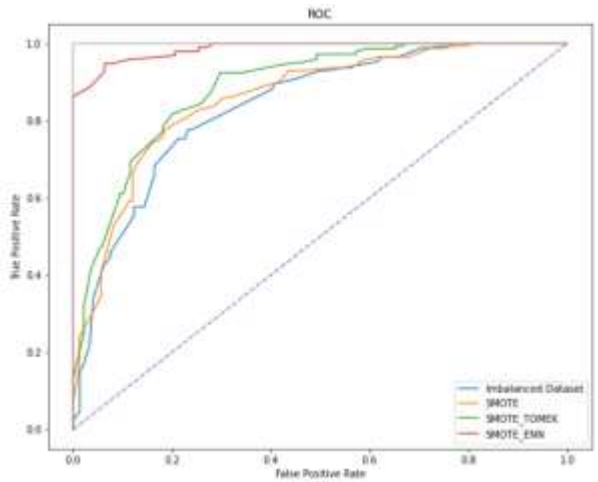

Fig. 7: ROC for random forest classification wine quality dataset

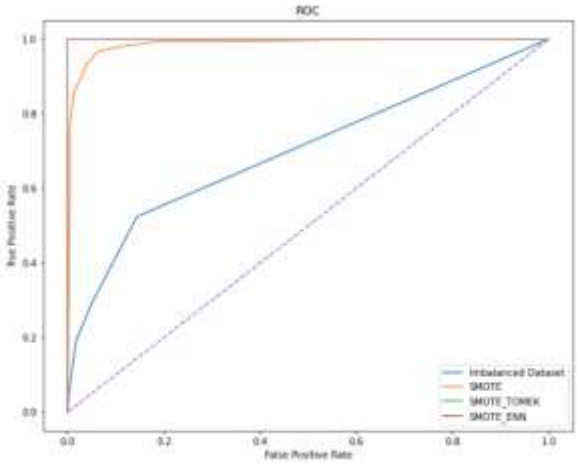

Fig. 8: ROC for AdaBoost classification diabetes dataset

Table 4: Performance results of the diabetes dataset

Classifier	Balancing technique		Precision	Recall	F-measure	Accuracy
AdaBoost classifier	Original (imbalanced)	0	0.78	0.88	0.83	0.7662
		1	0.73	0.58	0.64	
	SMOTE (balanced)	0	0.77	0.73	0.75	0.7433
		1	0.72	0.76	0.74	
	SMOTE Tomek (balanced)	0	0.80	0.81	0.81	0.8076
		1	0.81	0.81	0.81	
	SMOTE ENN (balanced)	0	0.98	0.95	0.97	0.9747
		1	0.97	0.99	0.98	
Random classifier	Original (imbalanced)	0	0.77	0.92	0.84	0.7792
		1	0.79	0.54	0.64	
	SMOTE (balanced)	0	0.78	0.78	0.78	0.7733
		1	0.76	0.76	0.76	
	SMOTE Tomek (balanced)	0	0.99	0.99	0.99	0.9929
		1	0.99	0.99	0.99	
	SMOTE ENN (balanced)	0	1.00	1.00	1.00	1.0000
		1	1.00	1.00	1.00	

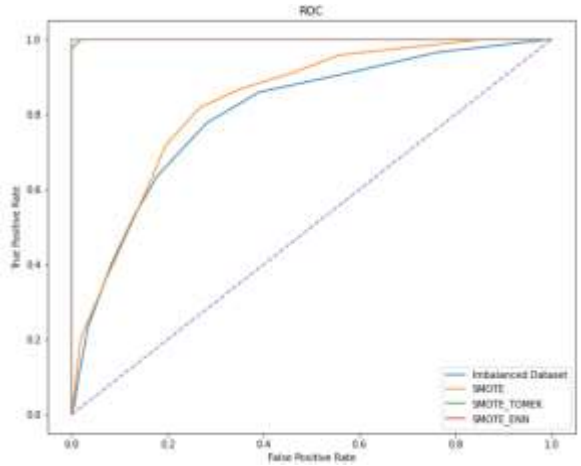

Fig. 9: ROC for random forest classification diabetes dataset

Diabetes Dataset

In AdaBoost classification, the unbalanced dataset and SMOTE have 0.7 accuracy. SMOTE Tomek has 0.8 accuracy. SMOTE ENN has a maximum of 0.9 accuracy. Whereas in random forest classification, SMOTE Tomek has got 0.9 and SMOTE ENN has 1.0 accuracy. This shows that SMOTE ENN is performing well when compared with all other methods for the diabetes dataset.

Through experiments, values for F-measure, accuracy, precision and recall are obtained. Figures 8-9 show that SMOTE ENN is performing well when compared with all other methods for the diabetes dataset. Table 4 contains a list of the computational results.

Discussion

The experiment is conducted for three datasets and results were compared with two classifiers and three state-of-the-art balancing algorithms.

The Classification accuracy of the AdaBoost classifier and random forest classifier are depicted in the bar chart shown in Fig. 10.

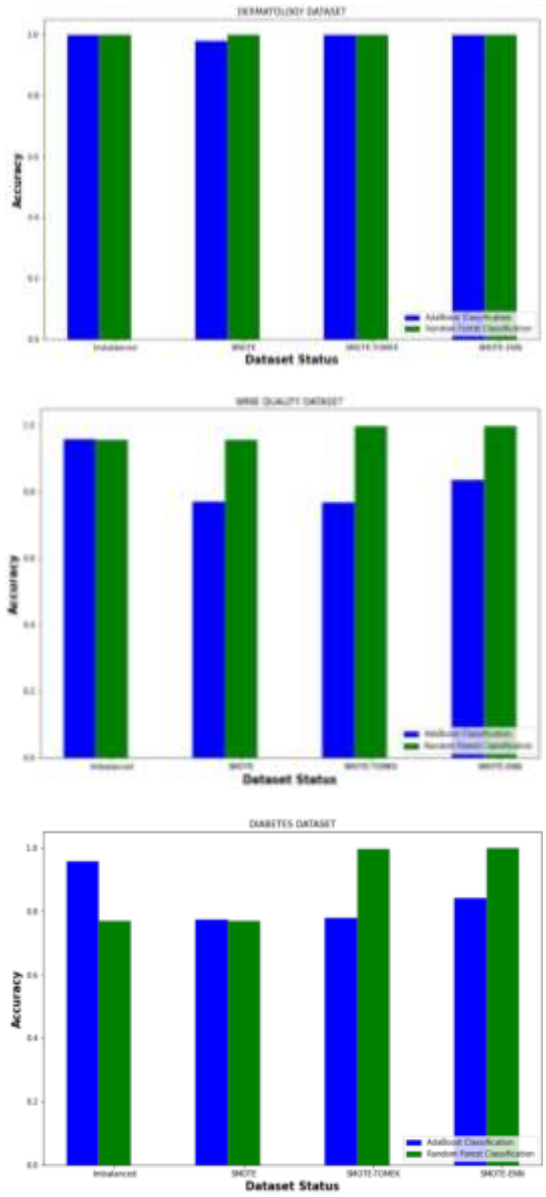

Fig. 10: Classification accuracy of AdaBoost and random forest classifier

With inference from the above bar chart, classification accuracy is 1.0 when applying the SMOTE ENN technique for balancing the dataset with the Random Forest classifier. SMOTE, SMOTE Tomek and SMOTE ENN balancing techniques are used to achieve the balancing, along with two classifiers: Random forest and AdaBoost. All three datasets produce 1.0 accuracy while using SMOTE ENN with random forest classification.

The classification results after applying pre-processing techniques are 1.0 in the random forest classifier and 0.9 in the AdaBoost classifier. When comparing pre-processing techniques, SMOTE ENN produces higher accuracy than SMOTE and SMOTE Tomek.

SMOTE ENN balancing technique has shown enhanced performance than the existing balancing techniques namely SMOTE and SMOTE Tomek. The result of the SMOTE ENN with Random forest classifier is compared with other existing techniques such as SMOTE and SMOTE Tomek.

Compared with the result of recent literature about SMOTE Tomek (Hairani *et al*., 2023) SMOTE ENN with random forest method provide high accuracy and precision for Diabetes dataset.

From the random forest classification and AdaBoost classification, we can identify that classification results are high in SMOTE ENN with the random forest classification method. With inferences from the above results, we conclude that the SMOTE ENN with Random forest classifier method is an optimized method for an imbalanced dataset.

In future, we can apply this method for multi-class imbalanced big data to balance dataset and improve the classification accuracy.

Conclusion

The inherent imbalance of many real-world problems across several classes has been addressed in recent years through the use of ensemble learning techniques (Tanha *et al*., 2020). Nevertheless, there haven't been enough studies in the literature to look at and contrast how well equalization algorithms perform with various classification techniques for this kind of dataset.

In this study, we tested state-of-the-art algorithms for preprocessing imbalanced data and compared their performance with two existing multiclass imbalanced data classification algorithms (Zhao *et al*., 2021). The experimental findings show that when the AdaBoost and random forest ensemble classifiers are applied to the three data sets using the SMOTE, SMOTE Tomek and SMOTE ENN methods, the SMOTE ENN method performs noticeably better than the other methods. It demonstrates its precision and accuracy.

A random forest classifier is a classifier optimized for classifying balanced datasets. Combined SMOTE ENN achieves high accuracy with low false negatives and low false positives. Therefore, an optimal ensemble method for balancing and classifying imbalanced datasets is SMOTE ENN using the random forest classifier technique.

In order to balance and pre-process the multi-class imbalanced dataset, a novel ensemble classification algorithm combining SMOTE ENN and random forest technique must be implemented. This can be developed to improve the classification accuracy level while requiring less time complexity for multi-class imbalanced big data streams.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Madhura Prabha R: Conception, designed, acquisition of data analysis, interpreted and drafted the article.

Sasikala S: Reviewed it critically and gave final approval.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

Competing Interests

The authors declare that they have no competing interests.

References

Amrehn, M., Mualla, F., Angelopoulou, E., Steidl, S., & Maier, A. (2018). The random forest classifier in WEKA: Discussion and new developments for imbalanced data. *arXiv preprint arXiv:1812.08102*.

An, T. K., & Kim, M. H. (2010, October). A new diverse AdaBoost classifier. In *2010 International Conference on Artificial Intelligence and Computational Intelligence* (Vol. *1*, pp. 359-363). IEEE. https://doi.org/10.1109/AICI.2010.82

Bader-El-Den, M., Teitei, E., & Perry, T. (2018). Biased random forest for dealing with the class imbalance problem. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(7), 2163-2172. https://doi.org/10.1109/TNNLS.2018.2878400

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20-29. https://doi.org/10.1145/1007730.1007735

Brzezinski, D., Stefanowski, J., Susmaga, R., & Szczech, I. (2019). On the dynamics of classification measures for imbalanced and streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(8), 2868-2878. https://doi.org/10.1109/TNNLS.2019.2899061

Çetin, A., & Gökhan, T. (2018). Differential diagnosis of erythematous squamous diseases with feature selection and classification algorithms. In *Nature-Inspired Intelligent Techniques for Solving Biomedical Engineering Problems* (pp. 103-129). IGI Global. https://doi.org/10.4018/978-1-5225-4769-3.ch005

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953

Clark, P., & Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Machine Learning-EWSL-91: European Working Session on Learning Porto, Portugal, March 6-8, 1991 Proceedings 5* (pp. 151-163). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0017011

Fernández, A., del Río, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced big data classification: Outcomes and challenges. *Complex and Intelligent Systems*, *3*, 105-120. https://doi.org/10.1007/s40747-017-0037-9

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets* (Vol. *10*, pp. 978-3). Cham: Springer. https://doi.org/10.1007/978-3-319-98074-4

Fernando, K. R. M., & Tsokos, C. P. (2021). Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 2940-2951. https://doi.org/10.1109/TNNLS.2020.3047335

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

Hairani, H., Anggrawan, A. & Priyanto, D., (2023). Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using SMOTE-Tomek Link. JOIV: International *Journal on Informatics Visualization*, *7*(1), pp.258-264.

Hasanin, T., & Khoshgoftaar, T. (2018, July). The effects of random undersampling with simulated class imbalance for big data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 70-79). IEEE. https://doi.org/10.1109/IRI.2018.00018

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Juez-Gil, M., Arnaiz-González, Á., Rodríguez, J. J., & García-Osorio, C. (2021). Experimental evaluation of ensemble classifiers for imbalance in big data. *Applied Soft Computing*, *108*, 107447. https://doi.org/10.1016/j.asoc.2021.107447

Junsomboon, N., & Phienthrakul, T. (2017, February). Combining over-sampling and under-sampling techniques for imbalance dataset. In *Proceedings of the 9th International Conference on Machine Learning and Computing* (pp. 243-247). https://doi.org/10.1145/3055635.3056643

Knerr, S., Personnaz, L., & Dreyfus, G. (1990). Single-layer learning revisited: A stepwise procedure for building and training a neural network. In *Neurocomputing: Algorithms, Architectures and Applications* (pp. 41-50). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-76153-9_5

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, *1*(2), 111-117. https://pzs.dstu.dp.ua/DataMining/preprocessing/bibl/Data-Preprocessing-for-Supervised-Leaning.pdf

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in Genetics*, *10*, 1077. https://doi.org/10.3389/fgene.2019.01077

Lin, W. J., & Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, *14*(1), 13-26. https://doi.org/10.1093/bib/bbs006

Lorena, A. C., De Carvalho, A. C., & Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, *30*, 19-37. https://doi.org/10.1007/s10462-009-9114-9

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, *7*, 93010-93022. https://doi.org/10.1109/ACCESS.2019.2927266

Ning, Q., Zhao, X., & Ma, Z. (2021). A novel method for Identification of Glutarylation sites combining Borderline-SMOTE with Tomek links technique in imbalanced data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(5), 2632-2641. https://doi.org/10.1109/TCBB.2021.3095482

Noorhalim, N., Ali, A., & Shamsuddin, S. M. (2019). Handling imbalanced ratio for class imbalance problem using SMOTE. In *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017) Transcending Boundaries, Embracing Multidisciplinary Diversities* (pp. 19-30). Springer Singapore. https://doi.org/10.1007/978-981-13-7279-7_3

Popel, M. H., Hasib, K. M., Habib, S. A., & Shah, F. M. (2018, December). A hybrid under-sampling method (HUSBoost) to classify imbalanced data. In *2018 21st International Conference of Computer and Information Technology (ICCIT)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICCITECHN.2018.8631915

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: A review. *International Journal of Computing and Business Research (IJCBR)*, *5*(4), 1-29. https://researchmanuscripts.com/July2014/2.pdf

Roshan, S. E., & Asadi, S. (2020). Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. *Engineering Applications of Artificial Intelligence*, *87*, 103319. https://doi.org/10.1016/j.engappai.2019.103319

Sleeman IV, W. C., & Krawczyk, B. (2021). Multi-class imbalanced big data classification on spark. *Knowledge-Based Systems*, *212*, 106598. https://doi.org/10.1016/j.knosys.2020.106598

Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets-A brief survey of the recent State of the Art. *Engineering Reports*, *3*(4), e12298. https://doi.org/10.1002/eng2.12298

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data*, *7*, 1-47. https://doi.org/10.1186/s40537-020-00349-y

Tarekegn, A. N., Giacobini, M., & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, *118*, 107965. https://doi.org/10.1016/j.patcog.2021.107965

Viola, P., & Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*, *4*(34-47), 4. https://www.researchgate.net/publication/215721846_Robust_Real-Time_Object_Detection

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, *94*, 88-104. https://doi.org/10.1016/j.knosys.2015.11.013

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... & Stoica, I. (2012). Resilient distributed datasets: A Fault-Tolerant abstraction for In-Memory cluster computing. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)* (pp. 15-28). https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia

Zhao, D., Wang, X., Mu, Y., & Wang, L. (2021). Experimental study and comparison of imbalance ensemble classifiers with dynamic selection strategy. *Entropy*, *23*(7), 822. https://doi.org/10.3390/e23070822

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J. & Ning, G., (2018). Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, 6, pp. 4641-4652. https://doi.org/10.1109/ACCESS.2018.2789428