Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 8, pp. 2141–2164
DOI : 10.47974/JIOS-1778

Ensemble learning based predictive modelling on a highly imbalanced multiclass data

Manka Vasti *
University School of Information, Communication and Technology
Guru Gobind Singh Indraprastha University
New Delhi 110078
India

and

Department of Computer Science and Engineering
School of Engineering and Sciences
GD Goenka University
Gurugram 122103
Haryana
India

Amita Dev †
Directorate of Training and Technical Education
Delhi 110034
India

Abstract

Class imbalance in the real-world datasets is a big challenge and the domains such as fraud detection, calamity occurrences, bankruptcy prediction etc. are prone to class imbalance due to the nature of occurrences of the events. In this paper, the detailed research using six ensemble machine learning techniques is applied to the undersampled, oversampled and the original dataset and the results are compared. The results of the research study indicates that amongst the applied six ensemble learners, the best learner is Random Forest algorithm (with entropy gain) implemented using ten-fold cross validation on the SMOTE oversampled dataset. 0.95 AUC and 0.8689 accuracy i.e. an increase of 4% in accuracy and substantial

* *E-mail:* **mankavasti@gmail.com (Corresponding Author)**
† *E-mail:* **amita_dev@hotmail.com**

increase in other performance indicators is observed as compared to the remaining five ensemble classifiers.

Subject Classification: *(2010) 86A15 (Seismology), 68T05 (Learning and Adaptive Systems), 62H30 (Classification and Discrimination), 62H30 (Cluster Analysis), 68T10 (Pattern recognition).*

Keywords: *Ensemble learning, Data augmentation, Earthquake prediction, Cluster based undersampling, SMOTE.*

1. Introduction

From the past one decade, the paramount increase in big data has called for efficient mechanisms to scan through the data and deduce patterns in the datasets. The accessibility of a gargantuan amount of data and the computational resources, along with the advanced techniques can foster more robust models for any use case. Given the abundance of datasets available across all domains, such as e-commerce, earthquakes, tsunamis, healthcare, finance, climate science, and social media, it is indeed crucial to extract patterns from them. A system based on clustering and classification, TPbEC2, is developed to classify tsunami alert as Alert and No Alert based on the dataset of activities of sea turtles using best ranked $MFK_{means}IC+$ FH.GBML algorithm with an accuracy of 94.21 % [14].

One of the most significant steps in accurate classification of data is the selection of optimal hyperparameters. A research study based on Grid Search, Bayesian Optimization, and Random Search applied to numerous types of datasets using 5-fold cross-validation extracted optimal hyperparameters for the ensemble classifier namely decision tree using AdaBoost [29].

Datasets can vary widely in size and type including structured, semi-structured, and unstructured data. Large datasets may come with a big challenge of class imbalance, wherein one class may have a large number of data samples called as the majority class and the other class has too few samples called as the minority class. Various domains such as disease diagnosis in healthcare, natural disasters such as tsunamis and earthquakes, weather monitoring, and anthropogenic effects such as violent conflicts and financial fraud, email filtering and many more applications face the challenge of class imbalance in the datasets due to the nature of the occurrences of the events. For example, for a binary class dataset with imbalanced class ratio, one class may have more than 90% (approx.). of the dataset samples and the other class has just 10% (approx.). Imbalance in dataset leads to problems in preparing the Machine Learning

(ML) classifiers for the existing real-world applications. Kumar et al., [18] lists the different challenges faced due to imbalanced class distribution and the various machine learning algorithms used to evade the effects of class imbalance. In order to train an effective learner, it is essential to have a balanced class distribution otherwise the models may be overwhelmed by one class and may ignore the remaining class(es) especially in multiclass datasets. Earthquake prediction, which is based on the past history of occurrences of earthquakes majorly suffers from the consequences of class imbalance. There are several hundred shallow and low magnitude earthquakes occur in a year and the high magnitude deep earthquakes appear in rarity. This infrequency of higher magnitude earthquakes over lower magnitude earthquakes creates imbalances in the dataset. In order to overcome such challenges, several data augmentation techniques to oversample datasets are available such as Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN) etc. These techniques generate more samples synthetically for the minority classes to overcome the disparity in the number of datapoints. Sometimes, it becomes very difficult to handle the large amount of dataset due to limited availability of the computing resources available. Therefore, to overcome this limitation undersampling techniques such as Random Undersampling, Clustering based undersampling etc. may be pursued to reduce the sample set of the majority class and bring it equivalent to the minority class. Random Undersampling randomly deletes the samples from the majority class whereas Clustering based undersampling technique replaces the data samples with their cluster centroids [25]. The number of clusters are dependent on the number of data samples of the minority class.

The oversampling and undersampling techniques are performed to bring the balance in the skewed distribution. In this research paper, the experimental study is performed to predict the class from a multiclass distribution. The paper presents the implementation of six ensemble learning techniques using tenfold cross validation. The format of the paper includes Section-II to present the details of the dataset used, Section-III gives the related work in this domain, Section-IV lists the details of the evaluation metrics, Section-V provides the details of the methodology used, Section VI illustrates the detailed results, Section VII gives conclusion of the study and Section VIII gives list of references pursued for the study.

2. Dataset and Preprocessing

This section provides details of the real – world dataset used and the steps of pre-processing applied to the dataset in order to prepare it for the research study.

2.1 *Research Dataset*

In this paper the research is done on high-quality, large-scale STanford EArthquake Dataset (STEAD) - A Global Data Set of Seismic Signals for AI recorded by seismic instruments. The database consists of shallow crustal earthquakes that are recorded at the active tectonic regions of the world. It also comprises of deep earthquakes of hundreds of kilometres of depth. The original data set contains 32 attributes, 1.2 million samples, 19K hours recordings at "local" distances within 350 km of earthquakes by 2613 receivers, 450K earthquakes that happened from January 1984 till August 2018 in the United States and Europe [23]. The dataset is available in the waveform (.hdf5 format) occupying 91GB disk space and attribute data in .csv file format of 3.5 GB size at kaggle.com [26].

2.2 *Pre-processing*

The details of the STEAD dataset used are as shown in Table 1. The dataset contains 1.05 million earthquake samples ranging from shallow to deep and devastating earthquakes. Pre-processing of the dataset includes handling Null values, NaN values, removing redundant attributes and label encoding the target to distinguish 1 to 7 classes in the dataset. After pre-processing, a total of 8,28,094 earthquakes and 13 attributes are considered for the study. The attributes set comprising of time of the occurrence of the earthquakes, location i.e. latitude and longitude of the earthquake, attributes related to p-wave and s-waves, earthquake depth in kilometres, distance from the source in kilometres, azimuth degrees, coda end sample number and earthquake category. The highly imbalanced dataset is oversampled using SMOTE which produces 3.07 million earthquake instances approximately. The undersampling is applied using two methods namely random undersampling technique and clustering based technique. The random undersampling removes samples randomly from the majority class to reduce the class imbalance whereas clustering based undersampling generates 245 clusters with equal samples of majority and minority classes.

Table 1
Preprocessing applied on the dataset used for the prediction

Description	Quantity of Data Samples
Original Dataset (earthquake + noise)	1.2 million instances
Only earthquake dataset	10,50,000 instances
Total attributes in the original dataset	32
After applying preprocessing on the dataset: a) Handling Null and Nan values b) Removing redundant attributes c) Label encoding seven classes	8,28,094 instances and 13 attributes
Oversampled dataset size after applying SMOTE	30,65,251 instances and 13 attributes
Randomly Undersampled dataset / Clustering based undersampled dataset size	245 clusters

The research study in this paper involves analysing and comparing the results of the six ensemble learning techniques on the original dataset, oversampled dataset and undersampled dataset.

3. Related Work

Artificial Intelligence aims to replicate the human cognition capability to provide computational solutions to address the problems across different domains that were previously solved using the conventional statistical approaches. Machine Learning (ML) has the capability to train the machine without the explicit programming. Machine Learning (ML) can infer co-relation between the features of the dataset without prior knowledge about it.

ML enhanced earthquake prediction has evolved over time in analysing the real-time seismic data. The scope of earthquake data analysis includes finding the epicentral location of the fault, time prediction of future earthquakes, ground-motion prediction, classification into the type of earthquake category etc. This is required for the management of emergency hazard response spectra and post disaster activities. Due to the complexity of the earthquake events, it has always been challenging for the seismological experts to study all the aspects of the problem using their expertise and logic. Also, the uneven occurrences of earthquake events generate highly imbalanced dataset that makes the analysis all the more challenging. Therefore, more robust modelling is required for the

accurate prediction and analyses. Jain, Bajpai, and Pamula [16] used a modified DBSCAN for the anomaly detection in the time-series dataset using seasonality to find both local and global anomaly from the dataset. Cluster based undersampling approach on the imbalanced class distribution is used for improving the classification accuracy [25]. Class imbalance is the result of skewed distribution of data points in a dataset. Guo et al., [33] suggested four different levels to deal with the challenge of class imbalance i.e. changing class distributions by undersampling, oversampling and advanced sampling; features selection at the features level; manipulating classifiers at the classifiers level and ensemble learning for the final classification. Clustering based undersampling approach is applied in the study to identify the risk factors and predict future infections of the 4616 ICU patients. This type of undersampling helped build reliable classifiers to enhance accurate decision making as mentioned in [5]. The different versions of SMOTE enhanced the precision of the minority class. A few types include upsampling the minority class within the limits of the majority class whereas some are based on the usage of noise filters after applying SMOTE as mentioned in [27]. The extension of SMOTE that is based on the ensemble-based noise filter called SMOTE Iterative-Partitioning Filter (IPF) resampling method overcomes the challenges produced by noisy and borderline samples in the dataset [6]. Zhang and Mani [10] has used a case study on Near Miss undersampling method that deletes samples based on the distance to the other samples existing of the same class. The data samples of the majority class are segregated into clusters that are built on different training subsets. Each subset is created for all the minority class samples. The final output is derived by the aggregation of the results predicted from each subset. Kang, Cho and MacLachlan [17] observed improved modelling by experimenting with clustering undersampling applied on ensemble machine learning.

Banking frauds or bankruptcy prediction is an example of class imbalance problem due to the nature of occurrence of events. Incorrect prediction may lead to classifying the bankruptcy case into the nonbankruptcy class that is far more damaging than the average rate of classification accuracy. Therefore, effective classifiers need balanced datasets to produce accurate results. Taehoon and Hyunchul [11] used a hybrid approach for under-sampling comprising of a blend of the k-Reverse Nearest Neighbour (k-RNN) and One-Class Support Vector Machine (OCSVM) method for accurately classifying a data sample. Johnson and Khoshgoftaar [9] reviewed the experimental studies on applying deep learning in the presence of class imbalance. ContrastMiner [31] generates

a contrast vector basis the historical behaviour sequence of the consumer to distinguish fraudulent transactions from the genuine ones. The intrusion detection system suffers from highly imbalanced class datasets. RIPPER [2] works on balancing class distribution by using undersampling and oversampling techniques along with machine learning. Bauder, Khoshgoftaar, and Hasanin [22] worked on the effect of minority class classification in the field of fraud detection using the real-world Big Data using deep learning and thresholding [8]. A novel framework to handle highly imbalanced multiclass dataset is proposed in Hamid, Yusoff and Mohamed [13]. Bauder and Khoshgoftaar [21] considered multiclass distribution and analysed the fraud using various machine learning algorithms on the imbalanced dataset. A research study used fraud classification using ensemble learning and evaluated multiple metrics of various algorithms [7]. Research study for building an effective approach to handle problems raised by class imbalance is done for a rare cardiovascular disease using ensemble learning as mentioned in Liu, Wu and Li [12]. A novel heterogeneous ensemble method is proposed using K-nearest neighbour, Random Forest and Support Vector machine for the diagnosis of cardio vascular disease with reduced feature subset [1]. A novel approach for combating the imbalanced data distribution problem is proposed by applying bagging and boosting techniques [15]. Salunkhe and Mali [28] worked on the advantages of both data level and classifier ensemble approach in order to increase the classification efficacy. Shi and Zhang [19] proposed an ensemble tree classifier for classifying highly imbalanced dataset. A novel ensemble classification method designed to deal with imbalanced data by training each tree in the ensemble using synthetically generated balanced data [4]. Wang et al., [24] used linear interpolation based on the normal distribution of the dataset in order to generate more samples near the center of minority class and to avoid marginalization. Yang, Pan and Zhang [32] proposed SD-KMSMOTE to generate more samples of the minority class based on its spatial distribution. A mix of supervised machine learning and ensemble learning techniques were applied on the SEER dataset to generate classifier models using k-fold cross validation [3]. Rupapara et al., [30] used a blend of ADASYN and Chi-squared feature selection method in order to combat the challenges of class imbalance and high dimensionality of datasets. In order to increase the classification accuracy, the data imbalance ratio is increased by creating minority samples artificially [34]. Kulkarni, Revathy, and Patil [20] combines four boosting algorithms to generate variations of AdaC2 and CSB2.

4. Evaluation Metrics

To evaluate the performance or quality of the created Machine Learning model, various metrics such as accuracy, precision, recall, F1 score, Jaccard Score and Negative Log Loss are compared. These metrics, also known as performance metrics or evaluation metrics, helps to understand how well the model has performed on the given dataset. Therefore, the performance of the model can be increased by fine-tuning the hyper-parameters so that it can generalize well on the new or the unseen data.

4.1 *Accuracy*

The accuracy metric is determined as the number of correct predictions i.e. True Positive (TP) and True Negative (TN) to the total number of predictions i.e. sum of all True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) as given in (1).

$$Accuracy = (TP+TN) \ / \ (TP+FN+FP+TN) \tag{1}$$

4.2 *Precision*

The precision determines the number of correct positive prediction. It can be calculated as the TP to the total positive predictions as shown below in (2).

$$\text{Precision} = TP \ / \ (TP+FP) \tag{2}$$

4.3 *Recall or Sensitivity*

Recall or Sensitivity aims to find the ratio of actual positive that was identified incorrectly as shown below in (3).

$$\text{Re}call = TP / (TP+FN) \tag{3}$$

4.4 *F1 Score*

F1 Score can be calculated as the harmonic mean of Precision and Recall, assigning equal weight to each of them as shown in (4)

$$F1-score = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

4.5 *Area Under Curve*

Area Under Curve (AUC) evaluates the performance and gives an aggregate measure between 0 and 1. Therefore, an AUC equals to 0.0 prediction whereas an AUC with 100% correct predictions will have value equal to 1.0.

4.6 *Jaccard Similarity*

Jaccard similarity is denoted as the ratio of the number of correctly predicted values to number of wrongly remaining values of predicted and real as shown in (5). Higher value of the Jaccard index score denotes higher accuracy of the classifier.

$$Jaccard = (TP) / (TP + FP + FN) \tag{5}$$

4.7 *Logloss*

LoggLoss is used to denote the output of the classifier as the probability of the class predicted as shown in (6).

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(P_{ij}) \tag{6}$$

N = No. of rows in the test dataset
M = No. of fault delivery classes
Y_{ij} = 1 if the observation belong to class j, else 0
P_{ij} = Probability predicted that observation belong to class j.

5. Methodology

The present study is based upon the comparative results of six ensemble learning classifiers on the original, oversampled and undersampled dataset in python 3.x programming language. The ensemble classifiers used are Random Forest Classifier using 'entropy', Random Forest Classifier using 'gini', Random Forest Classifier using 'log_loss', Histogram based Gradient Boosting Classifier (HistGradientBoost), Extreme Gradient Boosting Classifier (XGBoost) and Voting Classifier. Figure 1 below illustrates the methodology in the simple form.

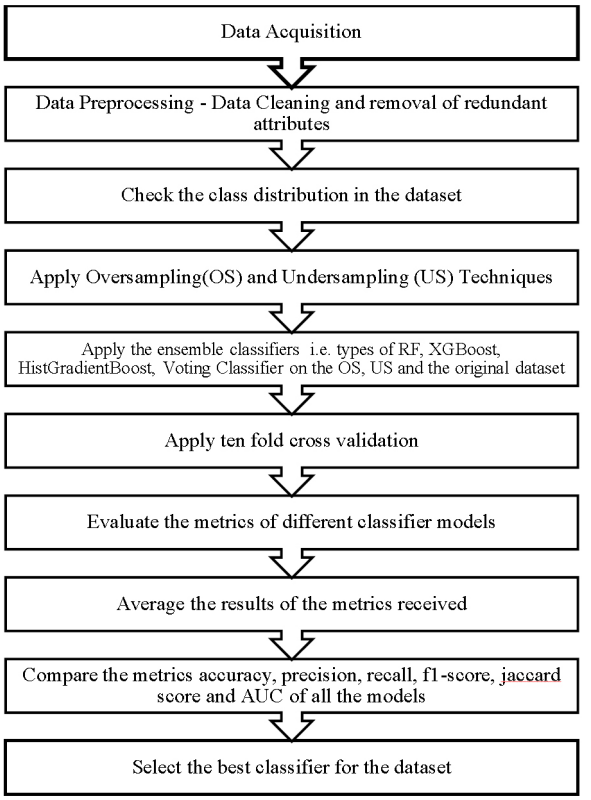

Figure 1

Block Diagram of the Methodology

After the acquisition of the gargantuan dataset, the first step is data pre-processing which includes data cleaning, handling Null and Nan values, removing redundant attributes and label encode the target. Next step includes the implementation of the ensemble learning classifiers namely Random Forest Classifier using 'entropy', Random Forest Classifier using 'gini', Random Forest Classifier using 'log_loss', Histogram based Gradient Boosting Classifier (HistGradientBoost), Extreme Gradient Boosting Classifier (XGBoost) and Voting Classifier. The performance metrics of all the ensemble classifiers for the multiclass dataset are compared. The models are trained, tested and validated using ten-fold cross validation methodology. As a result, 30 models are generated for each classifier and the value of the each of the metrics such as accuracy, precision, recall, f1 score, jaccard score, area under curve and negative

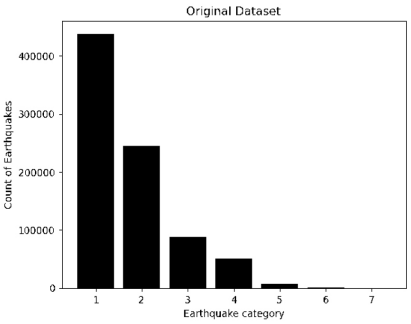

Figure 2

Class distribution of the original dataset

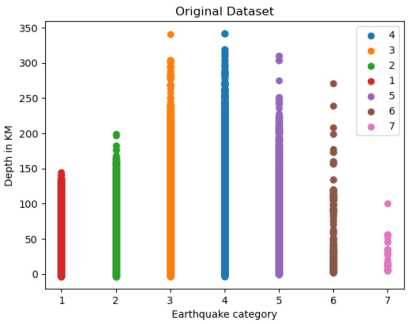

Figure 3

Plot between depth and earthquake category

log_loss is obtained by averaging their values. The best classifier is the one that provides highest accuracy and higher values of the performance metrics. The voting classifier compares multiple models that are trained, tested and validated together on the same folds of the data. The decision is based on soft voting where all the results are averaged to classify a data sample to one of the multiple classes.

STEAD dataset is highly imbalanced due to the nature of occurrence of earthquake events. Figure 2 illustrates the rarity of the occurrence of the higher magnitude earthquakes leading to imbalanced class distribution. The dataset is divided into multiple classes based on their magnitude and hence, this enlarges the problem statement to predicting the class of an earthquake in a multiclass environment.

The plot between the depth and magnitude category in the original dataset is as shown in Figure 3.

Since it is evident from Figure 2 that the dataset is highly imbalanced, it becomes extremely difficult to build an effective machine learning model for the earthquake prediction. Therefore, the following techniques are implemented in the research study to handle the class imbalance:

- Oversampling the dataset using Synthetic Minority Oversampling Technique (SMOTE)
- Undersampling the dataset by randomly removing the instances
- Undersampling based on Clustering

Oversampling using Synthetic Minority Oversampling Technique (SMOTE), is used to oversample all classes except majority class. The data samples are generated synthetically along the line segments joining the k minority class nearest neighbors. The result is as shown in the Figure 4 below:

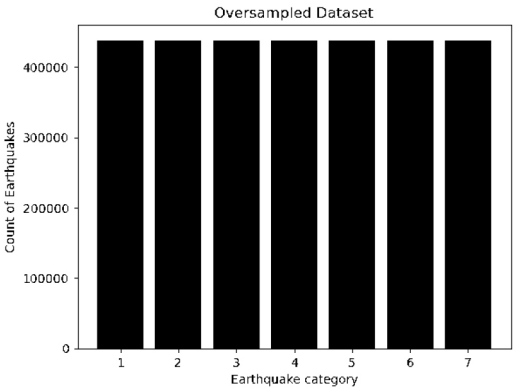

Figure 4

Oversampled Dataset

The clustering based undersampled dataset is as shown below in Figure 5:

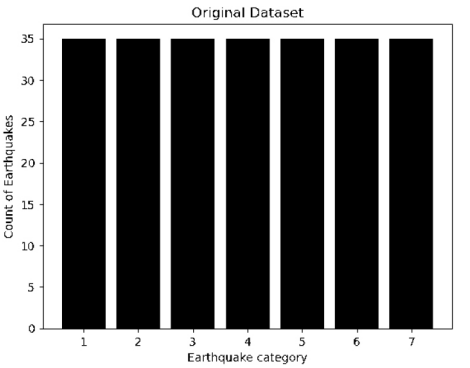

Figure 5

Undersampled Dataset based on Clustering

6. Results

The following section shows results of the comparative study conducted by implementing six ensemble learning techniques on STEAD

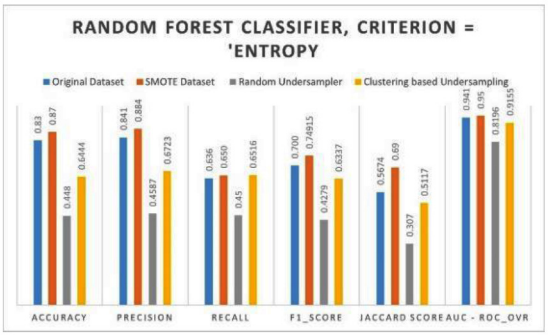

Figure 6

Random Forest, Criterion = Entropy

dataset. The experimental study is performed using tenfold cross validation generating 30 models in each of the six classifiers. The result metrics are then averaged and compared. Figure 6 below shows the comparison of the performance indicators when Random Forest with entropy gain, with 100 estimators is applied on the dataset. The results of this classifier are compared when applied on the original dataset, oversampled dataset and undersampled dataset. The six performance indicators of the classifier used for comparison include accuracy, precision, recall, F1 score, Jaccard score and AUC-ROC.

Table 2 shows the values of fine-tuned hyper-parameters such as number of estimators, value of k-fold, number of repeats, criterion applied and maximum features function. The scores of performance indicators when Random Forest (with 'entropy' criterion) is applied on the original, oversampled and undersampled dataset is as shown in Table 2.

Table 2

Random Forest (with entropy criterion), its parameters and values of the classifiers

Classifier	Parameters	Accuracy	Precision	Recall	F1_ Score	Jaccard Score	AUC_ ROC
Original Dataset	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.83	0.841	0.636	0.700	0.5674	0.941

Contd...

SMOTE Dataset	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.87	0.884	0.650	0.74915	0.69	0.95
Random Unders-ampler	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.448	0.4587	0.45	0.4279	0.307	0.8196
Cluster-ing based Unders-ampling	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.6444	0.6723	0.6516	0.6337	0.5117	0.9155

Figure 7 illustrates the details of the comparative results obtained after applying Random Forest (with 'Gini' criterion) and number of estimators are 100 to generate trees.

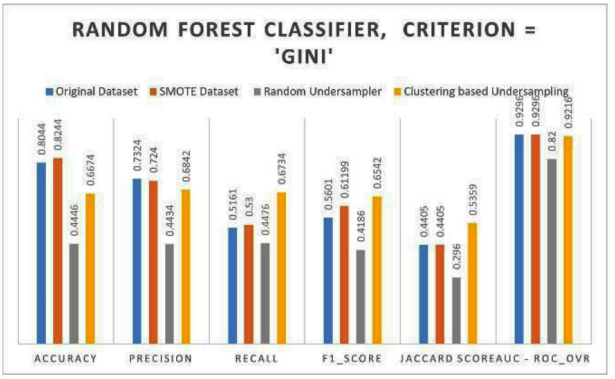

Figure 7

Random Forest, Criterion = Gini

Table 3

Random Forest (with gini criterion), its parameters and values of the classifiers

Classifier	Parameters	Accuracy	Precision	Recall	F1_ Score	Jaccard Score	AUC - ROC_ OVR
Original Dataset	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'gini', max_features = 'sqrt'	0.8044	0.7324	0.5161	0.5601	0.4405	0.9296
SMOTE Dataset	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'gini', max_features = 'sqrt'	0.8244	0.724	0.53	0.61199	0.4405	0.9296
Random Undersampler	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'gini', max_features = 'sqrt'	0.4446	0.4434	0.4476	0.4186	0.296	0.82
Clustering based Undersampling	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'gini', max_features = 'sqrt'	0.6674	0.6842	0.6734	0.6542	0.5359	0.9216

The scores of performance indicators when Random Forest, criterion = 'gini' is applied on the original, oversampled and undersampled dataset is as shown in Table 3.

The comparison of results of the metrics when Random Forest (with 'log_loss criterion) is applied on the original, undersampled and oversampled dataset is as shown below in Figure 8.

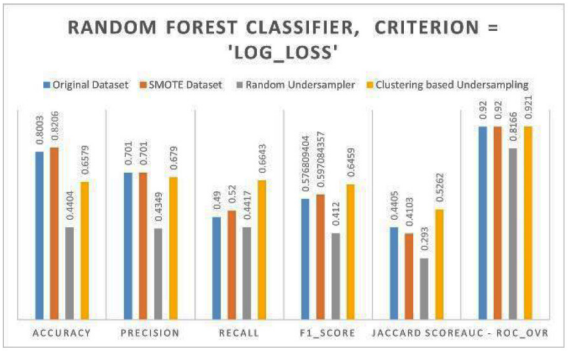

Figure 8

Random Forest, Criterion = log_loss

Table 4 indicates the scores of performance indicators when Random Forest, criterion = 'log_loss' is applied on the original, oversampled and undersampled dataset.

Table 4

Random Forest (with log_loss criterion), its parameters and values of the classifiers

Classifier	Parameters	Accuracy	Precision	Recall	F1_Score	Jaccard Score	AUC - ROC_OVR
Original Dataset	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='log_loss', max_features='sqrt'	0.8003	0.701	0.49	0.5768094	0.4405	0.92
SMOTE Dataset	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='log_loss', max_features='sqrt'	0.8206	0.701	0.52	0.59708436	0.4103	0.92

Contd...

Random Undersampler	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='log_loss', max_features='sqrt'	0.4404	0.4349	0.4417	0.412	0.293	0.8166
Clustering based Undersampling	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='log_loss', max_features='sqrt'	0.6579	0.679	0.6643	0.6459	0.5262	0.921

The comparison of working of Histogram Gradient Boosting classifier on the original, undersampled and oversampled dataset is as shown below in Figure 9.

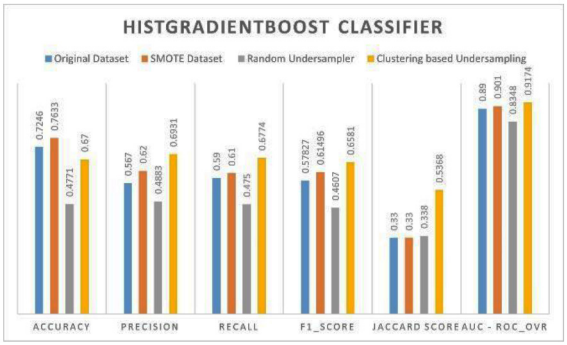

Figure 9

Histgradient Boost Classifier

Table 5

HistGradientBoost and its parameters and values of the classifiers

Classifier	Parameters	Accuracy	Precision	Recall	F1_Score	Jaccard Score	AUC - ROC_OVR
Original Dataset	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='entropy', max_features='sqrt'	0.7246	0.567	0.59	0.57827	0.33	0.89

Contd...

SMOTE Dataset	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='entropy', max_features='sqrt'	0.7633	0.62	0.61	0.61496	0.33	0.901
Random Undersampler	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='entropy', max_features='sqrt'	0.4771	0.4883	0.475	0.4607	0.338	0.8348
Clustering based Undersampling	n_estimators=100, n_splits(K Fold)=10, n_repeats = 3, criterion='entropy', max_features='sqrt'	0.67	0.6931	0.6774	0.6581	0.5368	0.9174

The results obtained by applying Histogram Gradient Boost Classifier along with the parameters are as shown below in Table 5.

The work done using XG Boost classifier is as shown below in Figure 10 and its parameters details are as shown in Table 6.

Table 6

Results from XGBoost Classifier

Classifier	**Parameters**	**Accuracy**	**Precision**	**Recall**	**F1_ Score**	**Jaccard Score**	**AUC - ROC_ OVR**
Original Dataset	n_estimators = 100, n_splits (K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.7546	0.671	0.671	0.4371	0.3335	0.8994
SMOTE Dataset	n_estimators = 100, n_splits(K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.7546	0.671	0.671	0.4371	0.3335	0.8994
Random Undersampler	n_estimators = 100, n_splits(K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.4852	0.4963	0.4837	0.4677	0.3382	0.835

Contd...

Clustering based Undersampling	n_estimators = 100, n_splits(K Fold) = 10, n_repeats = 3, criterion = 'entropy', max_features = 'sqrt'	0.68679	0.7256	0.6929	0.6805	0.5608	0.9081

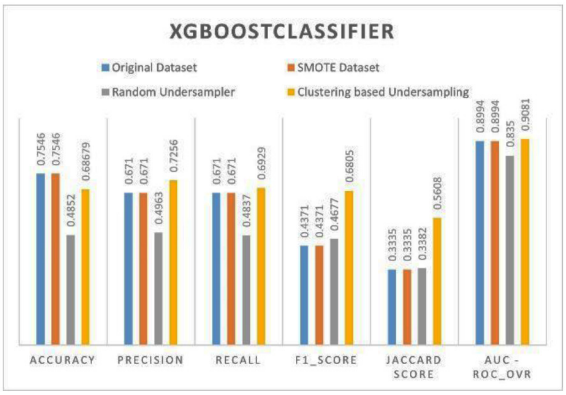

Figure 10

XGBoost Classifier

The mixed classifier model using Voting Classifier is made using Random Forest with entropy criterion, XGBoost and HistgradientBoost classifiers. Soft Voting is used and the results are averaged as shown below in Figure 11:

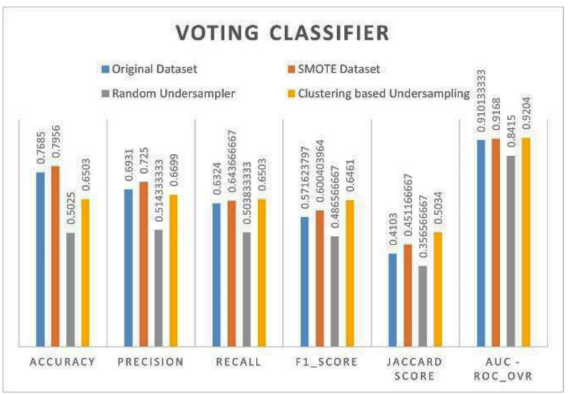

Figure 11

Voting Classifier

7. Conclusion

The class imbalance is the most common challenge in the world of growing data. In order to effectively train a learner, it is essential to have a balanced class distribution otherwise the models may be overwhelmed by one class and may ignore the remaining class(es) especially in multiclass datasets. The most related research studies have focused on applying either random undersampling or oversampling techniques using the typical train-test split methodology to train the model that produces only one model per classifier. The results obtained in such scenario is based on the split criterion. In this paper, the experimental study is conducted using one oversampling technique called SMOTE and two undersampling techniques namely random undersampling and clustering based undersampling. The comparative study is conducted based on the six ensemble learning techniques applied on the ten-fold cross validation with three repeats in each and hence generating 30 models in each classifier. The models are created using ten-fold cross validation for highest accuracy. The result metrics obtained from 30 models are averaged. The implementation of six ensemble learners is done on the original dataset, oversampled dataset and undersampled datasets generated. The first experimental study is done on the original dataset with the imbalanced class distribution. The various ensemble learning methodologies are applied on the original dataset include Random Forest Classifier using 'entropy', 'gini' and 'log_loss' criteria, HistGradientBoost Classifer, XGBoost Classifier and Voting Classifier. The second study used oversampled dataset and produced more than 3 million earthquake data samples synthetically by using Synthetic Minority Oversampling Technique (SMOTE). The same set of ensemble classifiers are then implemented on the oversampled dataset and the performance metrics are compared. Similarly, the third experimental study applied the ensemble classifiers on the undersampled datasets. Random undersampling and Cluster based undersampling were applied. As it can be seen from the results section, the best ensemble classifier has turned out to be Random Forest with entropy criterion formed using 100 estimators on the oversampled dataset of more than 3 million data samples as compared to the remaining 5 classifiers on the original, oversampled and undersampled datasets. It was observed that when Random Forest classifier with entropy gain and 100 estimators is applied using ten-fold cross validation on the multiclass SMOTE Oversampled dataset, a 4% increase in accuracy and substantial increase in other performance

indicators is observed as compared to other classifiers. This study will prove to be useful for the learners who want to pursue their research work in the domain of data augmentation.

References

[1] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset", *Computer Methods and Programs in Biomedicine,* vol. 198, (2021). doi: https://doi.org/10.1016/j.cmpb.2020.105770.

[2] D.A. Cieslak, N.V. Chawla, A. Striegel, "Combating Imbalance in Network Intrusion Datasets", in Proceedings of the International Conference on Granular Computing, Atlanta, USA, 10-12, pp. 732-737, May (2006). doi: 10.1109/GRC.2006.1635905.

[3] F. Deng, J. Huang, X. Yuan, C. Cheng and L. Zheng, "Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data", *Laboratory Investigation*, vol. 101, pp. 430–441 (2021), doi: https://doi.org/10.1038/s41374-020-00525-x.

[4] F. Kamalov, S. Moussa and J. A. Reyes, "KDE-Based Ensemble Learning for Imbalanced Data", Electronics 2022, vol. 11, Issue 17 (2022), doi: https://doi.org/10.3390/electronics11172703.

[5] F.S. Hernández, J.C. Ballesteros-Herráez, M.S. Kraiem, M. Sánchez-Barba, M.N. Moreno-García, "Predictive Modeling of ICU Healthcare-Associated Infections from Imbalanced Data. Using Ensembles and a Clustering-Based Undersampling Approach", Applied Sciences, vol. 9 (2019), doi:10.3390/app9245287.

[6] J.A. Sáez, J. Luengo, J. Stefanowski and F.Herrera, "SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering", Information Sciences, vol. 291, pp. 184–203 (2015), doi:https://doi.org/10.1016/j.ins.2014.08.051.

[7] J.M. Johnson and T.M. Khoshgoftaar, "Data-Centric AI for Healthcare Fraud Detection", SN Computer Science, vol. 4, no. 4, (2023), doi: 10.1007/s42979-023-01809-x.

[8] J.M. Johnson and T.M. Khoshgoftaar, "Deep Learning and Thresholding with Class-Imbalanced Big Data", in Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 755-762 (2019).

[9] J.M. Johnson and T.M. Khoshgoftaar, "Survey on deep learning with class imbalance", *Journal Big Data*, vol. 6, no. 1, (2019), doi:https://doi.org/10.1186/s40537-019-0192-5.

[10] J.P. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction", in Proceedings of the International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets, Washington, DC, USA, 21 August (2003).

[11] K. Taehoon and A. Hyunchul, "A hybrid under-sampling approach for better bankruptcy prediction", *Journal of Intelligent Information Systems*, vol. 21, no. 2, pp.173–190 (2015), doi: http://dx.doi.org/10.13088/jiis.2015.21.2.173.

[12] L. Liu, X. Wu and S. Li, "Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection", BMC Medical Informatics and Decision Making, vol. 82. https://doi.org/10.1186/s12911-022-01821-w

[13] M. H. A. Hamid, M. Yusoff and A. H. Mohamed, "Survey on Highly Imbalanced Multi-class Data", *International Journal of Advanced Computer Science and Applications* (2022), doi:10.14569/ijacsa.2022.0130627.

[14] N. Jain and D. Virmani, "Ensemble learning using fast rule based fuzzy K-means pre clustering and classification for aquatic behavior-extracted tsunami prediction," *Journal of Information and Optimization Sciences*, vol. 40, no. 2, pp. 441-453 (2019). doi: 10.1080/02522667.2019.1580884.

[15] P. Chujai, K. Chomboon, P. Teerarassamee, N. Kerdprasop and K. Kerdprasop, "Ensemble Learning For Imbalanced Data Classification Problem", in proceedings of International Conference on Industrial Application Engineering, January (2015), DOI: 10.12792/iciae2015.079.

[16] P. Jain, M. S. Bajpai and R. Pamula, "A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality", *The International Arab Journal of Information Technology* (IAJIT), vol. 19, no. 1, pp. 23 -28, January (2022), doi: 10.34028/iajit/19/1/3.

[17] P. Kang, S. Cho and D.L. MacLachlan, "Improved response modeling based on clustering, under-sampling, and ensemble", Expert Systems Applications, vol. 39, pp.6738–6753 (2012), doi:10.1016/j.eswa.2011.12.028.

[18] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data: Review of Methods and Applications", IOP Conf. Series: Materials Science and Engineering, vol. 1099, no.1, March (2021), IOP Publishing, doi:10.1088/1757-899X/1099/1/012077.

[19] P. Shi and Z. Wang, "An Ensemble Tree Classifier for Highly Imbalanced Data Classification", *Journal of Systems Science and Complexity*, vol. 34, no. 6, pp.2250–2266, 2021. doi:https://doi.org/10.1007/s11424-021-1038-8

[20] R. Kulkarni, S. Revathy, and S. Patil, "A Novel Approach to Maximize G-mean in Nonstationary Data with Recurrent Imbalance Shifts", *The International Arab Journal of Information Technology* (IAJIT), vol. 18, no. 1, pp. 103 - 113, January (2021), doi: 10.34028/iajit/18/1/12.

[21] R.A. Bauder and T.M.Khoshgoftaar, "The Effects of Varying Class Distribution on Learner Behavior for Medicare Fraud Detection with Imbalanced Big Data", *Health Information Science and Systems*, vol. 6, no. 1, (2018), doi: 10.1007/s13755-018-0051-3.

[22] R.A. Bauder, T.M. Khoshgoftaar and T. Hasanin, "An Empirical Study on Class Rarity in Big Data", in Proceedings of the 17th IEEE international conference on machine learning and applications (ICMLA), pp. 785–790 (2018), https ://doi.org/10.1109/ICMLA.2018.00125.

[23] S. M. Mousavi, Y. Sheng, W. Zhu and G. C. Beroza, "STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI," in IEEE Access, vol. 7, pp. 179464-179476 (2019), doi: 10.1109/ACCESS.2019.2947848.

[24] S. Wang, Y. Dai, J. Shen, L. Zhang, Y. Wang, and C. Zhang, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," Sci. Rep., vol. 11, no. 24039 (2021), doi: 10.1038/s41598-021-03430-5.

[25] S. Yen and Y. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions", Expert Systems with Applications, vol. 36, no. 3, Part 1, pp.5718-5727 (2009), ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2008.06.108.

[26] Stanford Earthquake Dataset (STEAD)- A Global Data Set of Seismic Signals for AI, IEEE Access, September 2021.[online]. Available: https://www.kaggle.com/datasets/isevilla/stanford-earthquake-dataset-stead/data.

[27] T. Maciejewski and J. Stefanowski, "Local neighbourhood extension of SMOTE for mining imbalanced data," in Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, Paris, France, pp. 104–111 (2011), doi: 10.1109/CIDM.2011.5949434.

[28] U. R. Salunkhe and S. N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach", Procedia Computer Science, vol. 85, pp. 725-732 (2016), ISSN 1877-0509,doi: https://doi.org/10.1016/j.procs.2016.05.259.

[29] V. J. Kadam and S. M. Jadhav, "Performance analysis of hyperparameter optimization methods for ensemble learning with small and medium sized medical datasets," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 1, pp. 115-123 (2020). doi: 10.1080/09720529.2020.1721871.

[30] V. Rupapara, F. Rustam, W. Aljedaani, H. F. Shahzad, and E. Lee, "Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model," Scientific Reports, vol. 12, no. 1000 (2022), doi: https://doi.org/10.1038/s41598-022-04835-6.

[31] W. Wei, J. Li, L. Cao, Y. Ou and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data", World Wide Web, vol. 16, no. 4, pp.449–75 (2013), doi:https://doi.org/10.1007/s11280-012-0178-0.

[32] W. Yang, C. Pan, and Y. Zhang, "An oversampling method for imbalanced data based on spatial distribution of minority samples SDKMSMOTE," Scientific Reports, vol. 12, p. 16820 (2022), doi: https://doi.org/10.1038/s41598-022-21046-1.

[33] X. Guo, Y. Yin, C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem", presented at the Fourth International Conference on Natural Computation, vol. 4, pp.192-201, 18-20 Oct. (2008), ISSN 978-0-7695-3304-9/08, DOI 10.1109/ICNC.2008.871.

[34] Y. Mi, "Imbalanced classification based on active learning SMOTE", *Research Journal of Applied Sciences, Engineering and Technology*, vol. 5, no. 3, pp. 944–949 (2013).

Received January, 2024