A Framework Using Binary Cross Entropy - Gradient Boost Hybrid Ensemble Classifier for Imbalanced Data Classification

S. Josephine Isabella
Research Scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Chennai, India. E-mail: josephineisabella@yahoo.co.in

Sujatha Srinivasan
Associate Professor, Department of Computer Science, SRM Institute for Training and Development, Chennai, India. E-mail: ashoksuja08@gmail.com

G. Suseendran
Assistant Professor, Department of Information Technology, Vels Institute of Science, Technology and Advanced Studies, Chennai, India. E-mail: suseendar_1234@yahoo.co.in

Received December 06, 2020; Accepted February 20, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18I1/WEB18076

Abstract

During the big data era, there is a continuous occurrence of developing the learning of imbalanced data gives a pathway for the research field along with data mining and machine learning concepts. In recent years, Big Data and Big Data Analytics having high eminence due to data exploration by many of the applications in real-time. Using machine learning will be a greater solution to solve the difficulties that occur when we learn the imbalanced data. Many real-world applications have to predict the solutions for highly imbalanced datasets with the imbalanced target variable. In most of the cases, the target variable assigns or having the least occurrences of the target values due to the sort of imbalances associated with things or events strongly applicable for the users who avail the solutions (for example, results of stock changes, fraud finding, network security, etc.). The expansion of the availability of data due to the rise of big data from the network systems such as security, internet transactions, finance manipulations, surveillance of CCTV or other devices makes the chance to the critical study of insufficient knowledge from the imbalance data when supporting the decision making processes. The data imbalance occurrence is a challenge to the research field. In recent trends, there is more data level and an algorithm level method is being upgraded constantly and leads to develop a new hybrid framework to solve this problem in classification. Classifying the imbalanced data is a challenging task in the field of big data analytics. This study mainly concentrates on the problem existing in most cases of real-world applications as an imbalance occurs in the data. This difficulty present due to the data distribution with skewed nature. We have analyses the data imbalance and find the solution.

This paper concentrates mainly on finding a better solution to this nature of the problem to be solved with the proposed framework using a hybrid ensemble classifier based on the Binary Cross-Entropy method as loss function along with the Gradient Boost Algorithm.

Keywords

Machine Learning, Imbalance, Preprocessing, Ensemble Classifier, Binary-cross-entropy Method, Sigmoid Function, Gradient Boost.

Introduction

The pattern recognition having an important task is classification. There are many classification algorithms such as decision trees, nearest neighbor, SVM, Logistic Regression that have been developed and applied successfully in numerous applications. But, the imbalanced nature of a data set will lead to concentrate on difficulties that occur during the learning process of the classifiers. In the class distribution, the occurrence of many more instances of a particular class than other classes is called an imbalanced dataset. Due to the presence of rare instances, the classification rules may predict to ignore or not discover the tiny classes[1]. Classification is the supervised learning method and is used to classify the new unknown data instances based on the generated classifier. Classification is used in various domains having their applications practically in the research field. There are more imbalanced datasets that occur in real-world applications[2]. The availability and growth of raw data are explored at the highest rate due to the developments of science and technology in recent years. This created an opportunity for gathering the information also known as knowledge discovery and the research with data engineering leads to a variety of applications from a low level to a high level. Recently, the class imbalance problem was raised from industry and academia domains. The basic issue of the class imbalance problem is that it has to adjust to the existing algorithms significantly[3]. In many real-world applications, the distribution of data skewed with the appearance of classes with much more frequent samples called the majority class and the rare occurrence of samples known as a minority class. But there is a possibility of gaining useful and important knowledge from minority class too. To overcome this difficulty, there is a necessity to create a machine, learning-based intelligent model. This study is known as imbalanced data learning. Such a development more widely explored from the past two decades. This can be achieved through improving the classifiers at the algorithm level and even in the preprocessing stage by applying the concept of balancing data through sampling methods[4].

This paper coordinated as follows,related works are discussed in Section 2. Section 3 dictates that finding a solution to the binary class imbalanced problem and the proposed algorithm Binary Cross Entropy-Gradient Boost Hybrid Ensemble Classifier (BCE-GBHEC) organized and building ofthe proposed models using BCE-GBHEC are discussedin Section 4. Section5shows the experimental results and discussions in detail. Finally, Section 6presents the conclusion and future work.

Related Works

The authors like Z. Wu.et al. proposed an ensemble algorithm for handling the imbalanced data of insurance business and was found to be more suitable in the recommendation of an insurance product or customer analysis than the traditional classifiers SVM, Logistic Regression, etc. They gave a pathway for the researchers to explore the proposed algorithm to improve the accuracy value[5].Wang et.al proposed a model for fraud detection in credit cards based on the training data set using the C4.5 algorithm as a base evaluator and solves the imbalance nature of the dataset by partitioning the dataset and clustering them as majority class and minority class using the nearest neighbor of each class center. [6].

Sohony et al. presented an ensemble method by keeping the best of both random forest and neural networks which predicted the new sample with high accuracy and confidence level. This method experimented with the European cardholders dataset, the imbalanced one. They used two types of classification methods namely Random forests and Feed-forward Neural Networks - 3 feed-forward neural networks and 2 random forest - ensemble methods to achieve their goals as to minimize the fraudulent sample's misclassification and also the normal sample's misclassification. They aggregated the results of individual classifiers and presented the final classifiers. They wanted to extend their future work is to improve the accuracy and handle the dataset with a text value[7].Huda et.al proposed an ensemble model for software defect prediction considering the real software class imbalance datasets from PROMISE Repository Software Engineering databases. Developing an accurate and fast system for fault detection as their future work[8]. Ren et al. proposed an oversampling technique with Entropy-based Wasserstein Generative Adversarial Network (EWGAN) to produce more minority samples of data in imbalanced learning. This method experimented with two highly imbalanced datasets namely vowel0 and page-blocks from keel repository [9].

Le et al. made an effort to solve the bankruptcy dataset with the occurrence of the class imbalance problem. The experiments resulted in that the RFCI framework produced better performance than the existing GMBoost algorithm. The future work will be concentrated on the investigation of finding the optimal bankruptcy model that finds the method to normalize the feature values and method for feature selection method and solving the class imbalance problem with the cost-sensitive method[10]. Wang et al.developed a hybrid model using XG-Boosting and Logistic Regression and is verified with the German Credit dataset to find the accuracy and compare the AUC value with other baseline models like SVM, Naive Bayes, Random Forest, XGBoost, LR and GLR. [11].

Finding a Solution to Binary Class Imbalanced Problem

1) Problem Definition and Objectives of the Study

The imbalanced data distributions lead to achieving an effective approximation of the given function y=f(x1, x2, …,xq) that the values of predictable values mapped into the target variable values based on the training set T_d, i=1 to n. The mapping function used is given in (1),

$$\Phi(y): y \rightarrow [1,0] \quad (1)$$

Where y is the target variable which assigns ‘1’ for minority class and ‘0’ for majority class. We can also use the threshold relevance which gives the frontier and the target variable above that are relevant.

The imbalanced classification problem is defined as the highly skewed data distribution of classes. i.e., the class has a highly unequal number of samples leads to an imbalanced classification nature. When a class having significantly fewer samples called a minority class and the other class having more samples called the majority class. This type of class imbalance distribution is known as a binary classification problem. In this study, we develop a framework for handling this type of datasets having binary classification[12].

Our contributions are given by,

i. The given samples to be checked for the missing values
ii. The verification of the imbalanced nature of the dataset to be carryout in order to find the number of minority class samples as well as majority class samples.
iii. The imbalance ratio has been found.

iv. The given samples to be preprocessed with the sampling techniques(random sampling and oversampling).

v. The main objective of the study is that finding a solution to the imbalanced classification problem by applying the proposed algorithm to an imbalanced dataset and the evaluation metrics would be calculated.

vi. After applying the existing algorithms to that imbalanced dataset and experimented with the various evaluation metrics and the results are analyzed.

2) Architecture Diagram of Proposed Model

In Figure 1, the architecture diagram of the newly developed model is shown. The proposed model the imbalanced data set for doing the classification. During the preprocessing step, it has to check the presence of null values and checking the occurrence of class imbalance. Finding the imbalance ratio leads to apply the oversampling technique to prepare the balanced dataset. Then the proposed model BCE-GBHEC will be generated to process the data and validating the model with 70% of training samples and 30% testing samples. Finally, we will get the evaluated results successfully.

3) Methodology - Binary Cross Entropy-Gradient Boost Hybrid Ensemble Classifier (BCE-GBHEC)

Boosting is the machine learning method that has to find a single strong prediction rule by combining the weak thumb rules which are produced by the base learning algorithm rather than each of the weak rules[13]. Gradient Boosting is a robust machine learning algorithm that gives an ensemble of decision trees in a stage-wise arrangement as the predictive model. This produces a new weak learner to get less information from the existing one. This algorithm has to optimize the loss to reduce the error. Usually, in the Gradient Boost technique, the optimization can be done by mean squared error loss, log loss, or cross-entropy loss for regression and classification problems. The proposed algorithm used binary cross-entropy as a loss function and is to be optimized[14].

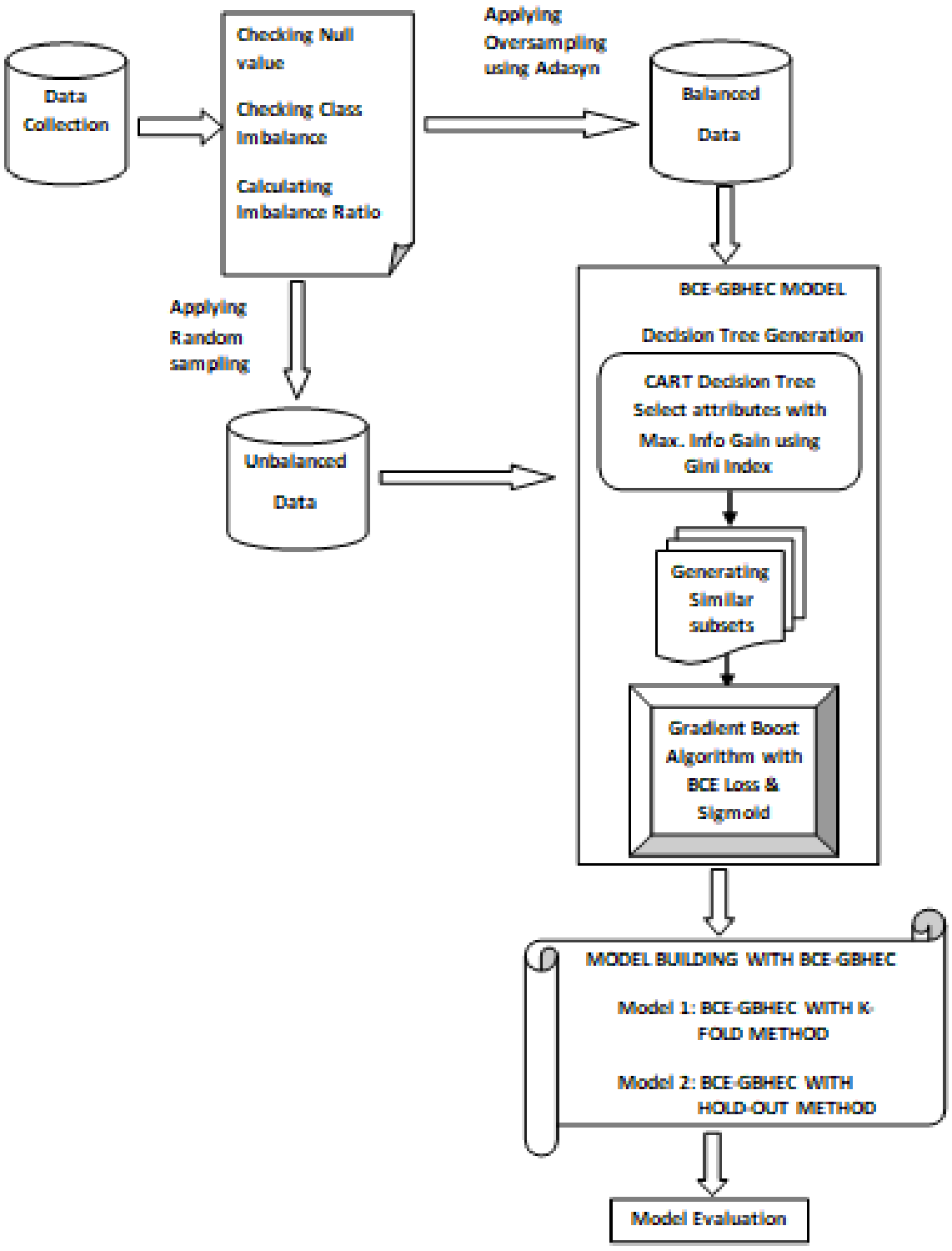

Figure 1 Architecture Diagram of BCE-GBHEC

Proposed Algorithm: BCE-GBHEC

The proposed algorithm BCE-GBHEC is defined as follows,

Input: Number of samples $s_1, s_2, s_3, \ldots . s_n$
Output: Improved Accuracy found in given 'n' samples
Begin
1. Collect $s_1(y), s_2(y), s_3(y) \ldots . s_n(y) \in R$
2. **For each sample** $s_i(y)$
3. Check for Null Values
4. Select the features
5. Construct Weak Learner $\{WL_1(y), WL_2(y), WL_3(y) \ldots . WL_n(y)\}$ using CART //Ensembled classifiers
6. Compute the split
7. Combine all Weak Learner as $X = \sum_{i=1}^{n} WL_i(y)$
8. **for each** $\boldsymbol{WL_i(y)}$
9. Compute the loss 'L' using the Binary Cross-Entropy method
   a).Compute the gradients
   b). Compute the predictive probabilities
   c) Fit the model
10. classify the samples as Fraud/Genuine based on the probabilities of the loss function
11. Detect the classifier with minimum loss $arg \min L[x, WL_i(y)]$
12. **End for**
13. Produce the Strong classifier
14. **End for**

End

1) Study of Proposed Algorithm

Gradient Boosting trains many models in a gradual, additive, and sequential manner. The proposed ensemble classifier uses the CART as a weak learner to construct a decision tree for classification.

CART is a statistical classifier to create binary trees by selecting the attributes and threshold that gives maximum information gain at each node of the binary tree [15]. The proposed method trains the weak learner to predict the samples based on the computation of negative gradients along with the learning rate to fit the new predictors.The ensemble classification process is illustrated in Figure2.

Let the number of samples as $s_1, s_2, s_3, \ldots, s_n$ as training data from $s_1(y), s_2(y), s_3(y), \ldots, s_n(y) \varepsilon R$.

For each sample,$s_i(y)$, we have to check the null values and select the features by using the CART.

Constructing the Weak Learners(WL) using CART as ensemble classifiers. They compute the split based on the greater information gain using the Gini index defined in (2),

$$Gini\ index = 1 - \sum_{i=1}^{n} {p_i}^2 \quad (2)$$

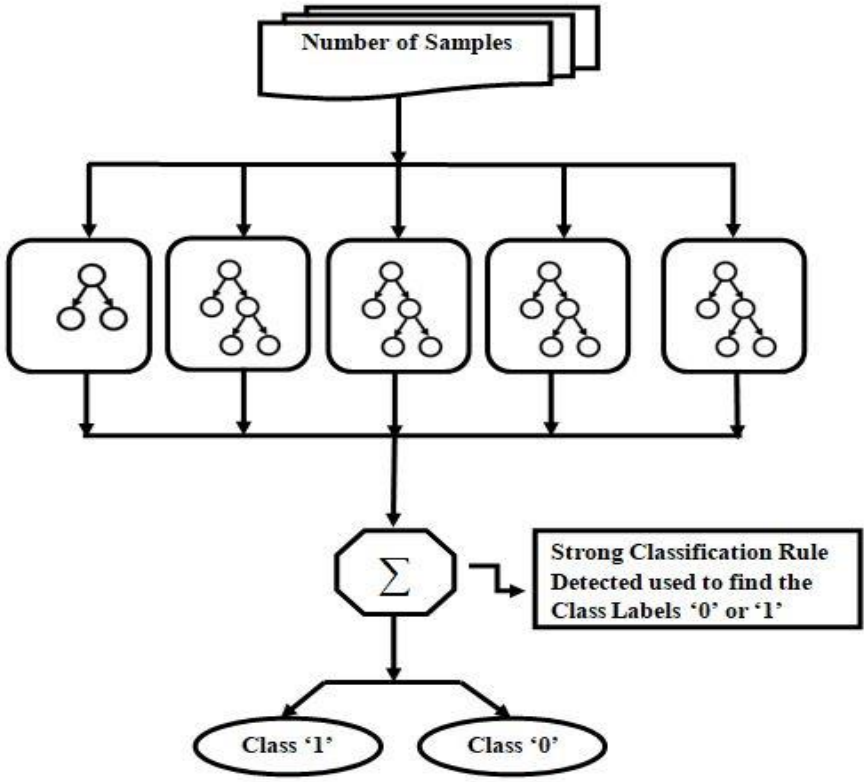

Figure 2 Ensemble classification

Then combine all Weak Learner as,

$$\sum \{WL_1(y), WL_2(y), WL_3(y) \dots\dots, WL_n(y)\} \quad (3)$$

Compute the Loss using Binary Cross-Entropy method by using, Eq.(4),

$$(y, \hat{y}) = \text{Minus}\frac{1}{N}\sum_{i=0}^{N} y * log(\hat{y}_i\ \{) + (1 - y)\ * log\ (1 - \hat{y}_i))\} (4)$$

where$\hat{y}$ is the predicted value and$y$ is the observed value.
Calculates the distance that how far away from the actual value (either '0' or '1').

Find the prediction is for each class and take the average of class-wise errors to give the final loss.

To find the optimization gradient of the loss function, the sigmoid function used Eq. (5).

$$s(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

Obtain the probabilities for the predicted values. Then fit the proposed model with the train and Test split of 70% - 30%.
Based on positive or negative probabilities of loss function predicts that,

$$y_i = 1 \Rightarrow \log\left(p(y_i)\right) \quad (6)$$

Classified as Positive classes – Class'1'

$$y_i = 0 \Rightarrow \log\left(1 - p(y_i)\right) \quad (7)$$

Classified as Negative classes- Class'2'
Finally, find the negative gradient of the classifier to fit the model as,

$$\arg\min L[x, WL_i\ (y)] \quad (8)$$

as steepest gradient.
This ensemble classifier framework, having a set of classifiers, each of which mapping to an instance vector y Ɛ R in the set of binary classes either '0' or '1'.
This ensemble does the primary tasks of constructing the classifiers individually and producing the classification rule that assigns the class label for y based on the results of the above classifiers.
The strong classification rule obtained is given in Table 1,

Table 1 Strong Classification Rule

```
If (attribute1=value-1) then Result=null // Terminal node
Else
If(attirbute2)= =value-2 and if(attribute3=value-3) then // Next node to check
…..
Else if(attribute n) =value-n then
Result = class '1' // Final Node that predicts the sample in class '1'
Else
Result = class '0' // Final Node that predicts the sample in class '0'.
End if
End if
```

2) Build the Proposed Models Using BCE-GBHEC

Both K-fold cross-validation and hold-out method are used to do the evaluation performance of the predictor in classification and regression[16].

a) Creating A Hybrid – Model Using BCE-GBHEC with K-Fold Method

k-Fold cross-validation is used as a standard method to evaluate the performance of the classifier. This method having an ability to split the given dataset into k divisions known as 'folds'. It uses k-1 divisions for training and one fold for the testing process. That is each of the instances getting the chance for individual testing. After the last iteration, we could consider the average performance measures for our research work[17][18]. Ahybrid model is created with BCE-GBHEC along with k-fold method to handle the imbalanced datasets.

b) Creating A Hybrid-Model Using Bce-Gbhec With Hold – Out Method

We have utilized the 'hold-out' method as the validation technique. This method splits the data into two mutually exclusive subsets. They are called as 'Train dataset' andTest dataset'[19]. This method training machine learning method by train data set and the model evaluation is carried out by the test dataset. Most of the universal applications use this method to find the evaluation measures to verify the performance of the applied classifier[17]. The proposed method BCE-GBHEC along with the k-fold method givesanother new model to classify the imbalanced data.

Experimental Results and Discussions

The proposed algorithm experiments with both the proposed models namely, hybrid-model using BCE-GBHEC with K-Fold Method and hybrid-model using BCE-GBHEC with Hold – Out Method.Themodel specification is defined in Table 2.

Table 2 Model Specification

- Proposed Algorithm: BCE-GBHEC-Binary Cross Entropy-Gradient Boost Hybrid Ensemble Classifier
- Base Learner: CART
- Loss Function Used: Binary Cross Entropy Loss
- Max-depth=3
- Treshold Value=0.5
- Rate at which the model learn=0.1
- Number of estimators=100
- Model build Validation checked: Number of Iterations as10- Iterationsm
- Number of Folds in K-Fold : 10 - Folds

1) Application of the Proposed Framework

Applying the proposed framework in "Fraud Detection System"(FDS). The fraud detection is an example of a classification problem of binary values with unbalanced data having more number of normal transactions(majority values) exceeds with fraudulent transactions(minority values)[20]. The FDS is an method which is used to identify the transactions which are fraudulent when the people made the transactions and must be intimated to the system administrator[21].

2) Dataset Description

The information about the dataset is described in Table 3.

Table 3 Dataset Detail

Name of the dataset	credit card fraud
Taken from	https://www.kaggle.com/mlg-ulb/creditcardfraud
Origin	Real-Time European cardholders in the year, September2013
Number of instances	2,84,807 instances
Number of attributes	30
Nature of the dataset	Binary classification-Imbalanced dataset(genuine class or Fraud class)

a) Checking for Class Imbalance

Streams of credit card transactions give that the classes are extremely unbalancedsince frauds are typically less than 1% of genuine transactions. The imbalancedataset becomes a balanced one by replicating the minority class (fraud cases) called oversampling. Figure 3 shows class imbalance clearly and thisimbalance leads to an incorrect prediction of the result due to the presence ofmore samples in the majority class.This imbalance can be solved usingoversampling with the Adasyn technique.

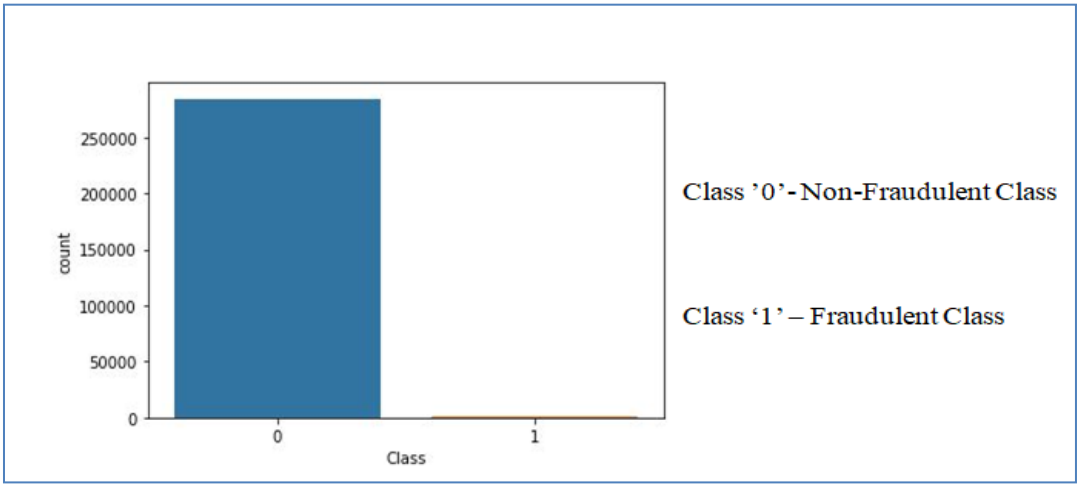

Figure 3 Class imbalance checking

b) Class Balancing - Oversampling Using Adasyn

Adasyn is used to generate minority samples based on their distributions adaptively that is, it easily generates synthetic data which is not easier to learn for the minimal class samples by shifting the decision margin to mainly aim at those specific hard to learn samples. This algorithm is suitable for the binary class imbalanced problem[22].

In Table4, the count of data samples of the original dataset and the oversampled samples are given. The imbalance ratio is defined by,

$$Imbalanc\ Ratio\ (IR) = \frac{Total\ count\ of\ Large\ Class\ Samples}{Total\ count\ of\ Minimal\ Class\ Samples} \quad (15)$$

The proportion between the total count of total samples in the superior class and the count of total samples in the minimal class[23].

Table 4 Original Dataset vs. Oversampled Dataset

Dataset	Total count of samples	Count of fraud samples	Count of genuine instances	Imbalance Ratio(IR)
Original Dataset	284807	492	284315	577.876
Oversampled using Adasyn	568555	284240	284315	1.000264

From the results of the imbalanced ratio (IR), the highly imbalanced data set becomes balanced by efficiently applying the ADASYN oversampling technique.

3) Results Obtained Using BCE-GBHEC with K-Fold Method

By applying the k-fold cross-validation method, our study took the consideration of k=10 folds to evaluate theBCE-GBHEC, and finally, we found that the performance accuracy for each fold and the final result produced with the average of 10-Folds. In random sampling, we got the average accuracy as 92.83% and having the performance timing as 2706.511 seconds. In the case of oversampled data, we found that the accuracy was 97.4931% and 2927.032 seconds as performance timing.

4) Results Obtained Using BCE-GBHEC with Hold – Out Method

We have applied the 'hold-out' method as the validation technique in this case of study. Usually, the standard split is to divide the entire dataset into the two-third portion of the data (approximated to 60% - 70%) as the train data and the remaining portion

(approximated to 30%-40%) as test data. Therefore we applied the hold-out method with a 70% -30% split of data[19] as validation. This experimentation produced 92.74% of accuracy and 283.494 seconds of performance timing for random sampling instances and the oversampling instances resulted in 97.23% of accuracy and 600.85 seconds of performance timing.

Table 5 K-Fold cross-validation vs. Hold out method

Validation Method	Sampling Type	Accuracy (%)	Performance Timing (in Seconds)
K-Fold Cross-Validation (Average of K=10)	Random Sampling	92.833745	2706.511
	Over Sampling	97.4931184	2927.032
Hold-Out Method (70%-30% of Split)	Random Sampling	92.74	283.494
	Over Sampling	97.22881	600.851

Table 5 shows the comparative status of both the k-fold and hold-out method along with random sampling instances and oversampling data. From the results, we observed that the k-fold method produced, the accuracy is more in case of oversampling and having a difference of 0.09% than random sampling data. But the performance timing is having the worst case. The oversampled data require more than 9 times than that of random sampling instances. Similarly, the hold-out method resulted in, the difference between the accuracy produced for random sampling data and oversampling data as 0.26%. The performance timing has the major difference as more than 4 times of performance timing is required by k-fold than that of the hold-out method. In both the sampling methods, the k-fold method having very mere improved variation in accuracy when compared with the hold out method.But the holdout method having a significant difference in performance timing than that of k-fold.Therefore we consider the hold-out method results for further experiments and doing the performance measures as well.

Initially, during the learning phase, the decision trees are more too sensitive to learn and in the training phase, they have disturbed due to the variance. To overcome these disturbances, the experiments were attained ten runs and the results are taken as average and used for the study[24]. The results obtained by running the model for 10 iterations to check the validation of the model and average values are obtained and are used to study with other classifiers, and finally, we have the accuracy as 97.22881 % ~ 97.23%.

5) Comparing Proposed Ensemble Classifier (BCE-GBHEC) vs. Other Classifiers

a) Accuracy Comparison

Accuracy is a common and most important measure of the predictive model that shows how the algorithm efficiently performed in overall studies.The accuracy values are tabulated in Table7.The proposed framework produces the highest accuracy of 97.23%. The KNN predicts the accuracy as 96.89% whereas Adaboost and AdaBoost and SVM produce 96.87% and 96.81% respectively. The graph is visualized in Figure 4.

Table 7 Results of Accuracy

Classifier	Accuracy
Adaboost	96.87%
KNN	96.89%
Adaboost+SVM	96.01%
BCE-GBHEC	97.23%

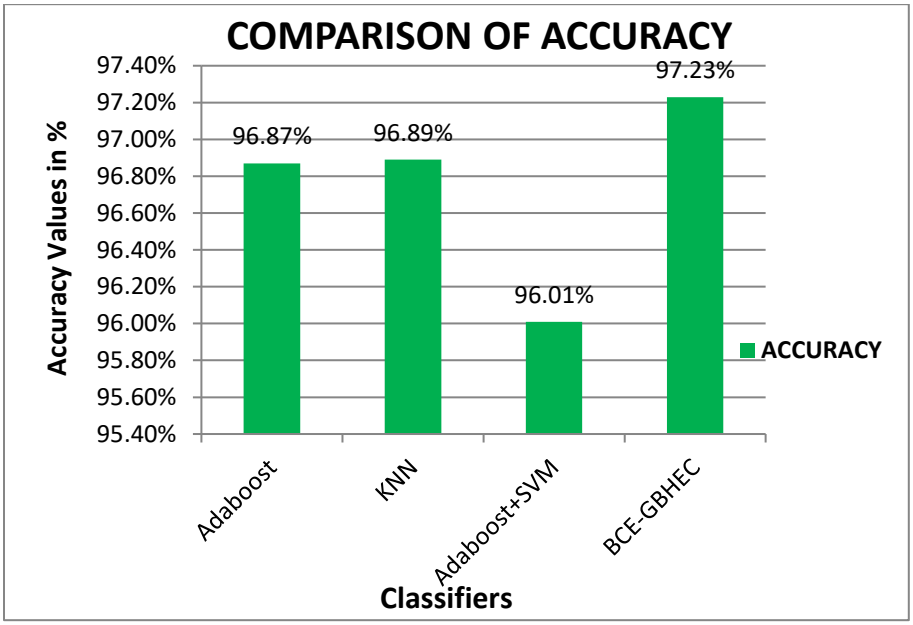

Figure 4 Comparison of Accuracy

The proposed classifier algorithm produces the 97.23%of accuracies improved 0.36% with Adaboost, 0.34%with KNN classifier and 1.22% that of Adaboost and SVM.

b) Comparison of Performance Timing of Proposed Model

When we classify the instances the time duration is the most important key factor and istaken for considerationin this study. The time duration for classification done by

different classifiers is given in Table 8. The proposed model has the time duration to dothe classification is 600.851 seconds.

Table 8 Comparison of Time Duration

Classifier	Time (in seconds)
Adaboost	605.952
KNN	604.357
Adaboost+SVM	602.952
BCE-GBHEC	600.851

The visualization of time duration is given in Figure5.

The proposed model having the least time duration when compared with other classifiers.

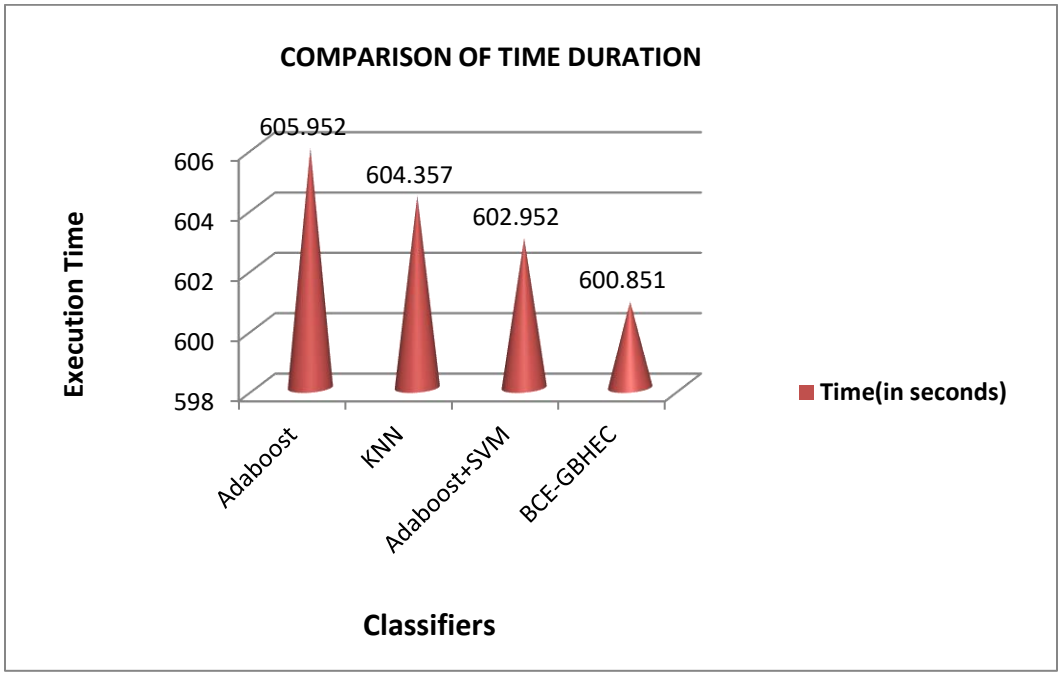

Figure 5 Comparison of Execution Time

The proposed model has the time duration as 5. 10 seconds less than that of Adaboostand 3.506 seconds less than that of the KNN classifier and 2.101 seconds lesser than Adaboost + SVM.

Conclusion and Future Work

In this paper, we have discussed that the more challenging problem occurred in the area of big data and big data analytics due to data exploration is the imbalanced data classification. Real-world problems such as medical diagnosis, customer retention, credit card fraud detection, churn prediction, and many more experience the class-imbalance.

This study found the impact of class imbalance and takes an opportunity to solve this problem with help of developing a hybrid ensemble classifier named BCE-GBHEC.We have learned a lot of lessons from the experiments of proposed work and make suggestions to do the possible things when we scrutinize the future work as well. This proposed framework has a limitation that it processes the binary class imbalance data set successfully and is suitable to find the results efficiently but not the multiclass imbalanced data.

References

Sun, Y., Wong, A.K., & Kamel, M.S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*(04), 687-719.

Le, T., Lee, M.Y., Park, J.R., & Baik, S.W. (2018). Oversampling techniques for bankruptcy prediction: novel features from a transaction dataset. *Symmetry*, *10*(4), 79.

Vluymans, S. (2019). Learning from imbalanced data. *In Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods*, *Springer, Cham*, *807*(9), 81-110.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

Wu, Z., Lin, W., Zhang, Z., Wen, A., & Lin, L. (2017). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE International Conference on Computer Science and engineering, IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017*, *1*, 531–536.

Wang, H., Zhu, P., Zou, X., & Qin, S. (2018). An ensemble learning framework for credit card fraud detection based on training set partitioning and clustering. *In IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 94-98.

Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. *ACM International Conference Proceedings Series*, 289–294.

Huda, S., Liu, K., Abdelrazek, M., Ibrahim, A., Alyahya, S., Al-Dossari, H., & Ahmad, S. (2018). An ensemble oversampling model for class imbalance problem in software defect prediction. *IEEE access*, *6*, 24184-24195.

Ren, J., Liu, Y., & Liu, J. (2019). EWGAN: Entropy-based Wasserstein GAN for imbalanced learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 10011-10012.

Le, T., Le Son, H., Vo, M.T., Lee, M.Y., & Baik, S.W. (2018). A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. *Symmetry*, *10*(7), 250.

Wang, M., Yu, J., & Ji, Z. (2018). Credit fraud risk detection based on XGBoost-LR hybrid model," *Proceedings of the International Conference on Electron Bus*, 336–343.

Zhao, Y., Wong, Z.S.Y., & Tsui, K.L. (2018). A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of healthcare engineering*, *2018*. https://doi.org/10.1155/2018/6275435

Schapire, R.E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, *Springer, New York, NY*, 149-171.

Niu, X., Wang, L., & Yang, X. (2019). A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv preprint arXiv:1904.10604*.

"Tree algorithms: ID3, C4.5, C5.0 and CART - Data Driven Investor - Medium." https://medium.com/datadriveninvestor/tree-algorithms-id3-c4-5-c5-0-and-cart-413387342164.

Baumgartl, H., Tomas, J., Buettner, R., & Merkel, M. (2020). A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring. *Progress in Additive Manufacturing*, 1-9.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, 515. https://doi.org/10.3389/fgene.2018.00515.

Wong, T.T., & Yang, N.Y. (2017). Dependency analysis of accuracy estimates in k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, *29*(11), 2417-2427.

Alam, M., Thapa, D., Lim, J.I., Cao, D., & Yao, X. (2017). Computer-aided classification of sickle cell retinopathy using quantitative features in optical coherence tomography angiography. *Biomedical optics express*, *8*(9), 4206-4216.

Padmaja, T.M., Dhulipalla, N., Krishna, P.R., Bapi, R.S., & Laha, A. (2007). An unbalanced data classification model using hybrid sampling technique for fraud detection. In *International Conference on Pattern Recognition and Machine Intelligence*, *Springer, Berlin, Heidelberg*, 341-348.

Isabella, S.J., Srinivasan, S., & Suseendran, G. (2020). An Efficient Study of Fraud Detection System Using Ml Techniques. *In Intelligent Computing and Innovation on Data Science*, *Springer, Singapore*, 59-67.

Dorn, K.H., & Jobst, T. (2002). Internal cleaning of steel piping systems. *JOT Journal for Surface Technology*, *42*(5), 56-57.

Liu, Z., Cao, W., Gao, Z., Bian, J., Chen, H., Chang, Y., & Liu, T.Y. (2020). Self-paced Ensemble for Highly Imbalanced Massive Data Classification. *In IEEE 36th International Conference on Data Engineering (ICDE)*, 841-852.

Gómez, S.E., Martínez, B.C., Sánchez-Esguevillas, A.J., & Callejo, L.H. (2017). Ensemble network traffic classification: Algorithm comparison and novel ensemble scheme proposal. *Computer Networks*, *127*, 68-80.