An Approach to Reliably Classify Imbalanced Data

Venkatesh Mainalli[#1], Shivraj Deshmukh[#2], Gaurav Bade[#3], Akshay Dhole[#4], Geeta S. Navale[*#5]

[#]Department of Computer Engineering, Sinhgad Institute of technology & Science, Savitribai Phule Pune University

[1]venkateshmainalli55555@gmail.com

[2]deshmukh.shivraj95@gmail.com

[3]gauravbade11@gmail.com

[4]akshaydholead10@gmail.com

[5]gsnavale_sits@sinhgad.edu

Abstract

Several real-world data sets have an imbalanced distribution of the instances. Learning from such data sets leads classifier being biased towards the majority class, thereby tending to misclassify the minority class samples. Imbalanced data set can cause negative effect on machine learning's classification performance. Many attempts are carried on for addressing issue of imbalanced datasets. The data is to be rebalanced by artificial means by oversampling or under sampling to handle the problem of imbalanced data. In this paper authors propose an approach referred as diversifying ensemble technique which can eliminate such drawback & the related work carried out in this domain.

Keywords—Imbalanced Dataset; classifiers; sampling; classifier ensemble

I. INTRODUCTION

An imbalanced dataset is exclusive case within which instances of one of the two classes is more than the opposite, in different way, the amount of observations isn't equivalent for all the classes in a classification dataset. A balanced data set is the one that contains equal or nearly equal range of samples from the positive and negative class. That dataset ought to be balanced with both the majority class and the minority class. Imbalanced dataset is common in several real-time problems from telecommunications, web, finance-world, ecology, biology, medicine, etc. which can be thought of one of the top problem in data processing these days. Moreover, it is worth to point out that the minority class is sometimes the one that has the highest interest from a learning point of view and it conjointly implies a big cost when it is not well classified.

A. Motivation

Multiple real-world data sets have an imbalanced distribution of the instances. Learning from such data sets leads classifier being skewed towards the class with majority count, thereby tending to misclassify the minority class samples. For example fraud detection, cancer detection, online advertisement conversion etc.

II. LITERATURE SURVEY

This section presents the related work carried out by the other researchers in the field of imbalanced data classification using various methods as illustrated in Fig. 1.

1. Data Level Approach
2. Algorithmic Level Approach

The Data Level Approach:

This approach includes sampling algorithms as follows:

1. Random Under-Sampling
2. Random Over-Sampling
3. Cluster Based Over Sampling
4. Informed Over-Sampling i.e. Synthetic Minority Over-Sampling Technique

The Algorithmic Level Approach:

This approach includes 3 types:

1. Threshold Moving Approach
2. Cost Sensitive Learning
3. Ensemble Learning

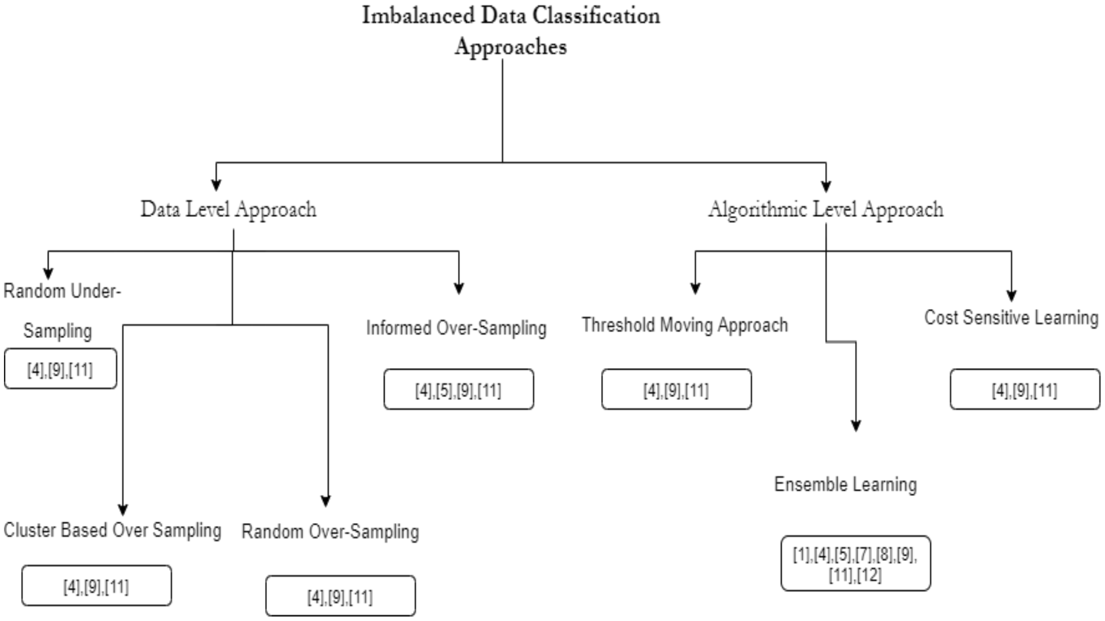


Fig. 1 Classification of algorithms to handle imbalanced dataset

A novel associative classification algorithm called Association Rule-based Classification for Imbalanced Datasets (ARCID) is introduced by Safa Abdellatif et. al[1]. ARCID is based on three stages which are generating, filtering and selecting rules. The first stage consists in generating frequent rules from each class of the training set using a local support. The second stage consists in filtering rules generated during the first stage. To do so, a new ranking and pruning technique is proposed supporting multiple criteria aggregation in order to keep simultaneously rules with a high predictive accuracy and those which are rare but of primary interest. The last stage consists in predicting the class label of the new data. Experimentations, against five real-world datasets obtained from the UCI repository and using different rule-based and non-rule-based approaches, have been conducted with reference to four assessment measures in order to evaluate the performance of the proposed approach. Techniques based on this approach have yield good accuracy compared to other classification techniques. However, mining imbalanced datasets was thought of as one of the top ten data mining challenges since most of the machine learning (ML) algorithms assume that datasets have balanced class distribution. However, managing the overwhelming number of Class Association Rules (CAR) generated from real-life datasets and Removing redundant rules conveying the same information is repetitive task.

By applying MapReduce paradigm to SplitBal algorithm, Jakub Neumann et al.[2] introduced two algorithms for classifying imbalanced dataset dissimilarity based imbalance data classification (EDBC) and splitting based data balancing method (SplitBal).In EDBC algorithm authors used three methods i.e. feature selection and data reduction, prototype selection and data transformation based on dissimilarity calculation. The basic idea of SplitBal is to divide the majority class instances into several bins so that each bin contains the minority class and a part of majority class of the equal size. Training of different bins is performed in parallel due to which no data exchange among computations for different bins is needed. However as EDBC isn't implemented in parallel, bigger dataset cannot be used for experiments.

T. Jaya Lakshmi et. al [3] detected many problems which occurred in real world such as fraud credit card transaction, medical diagnosis and email foldering due to the misclassification of the minority classes. They observed that the algorithms for imbalanced dataset cannot address the issue of imbalance data efficiently. The proposed algorithm treats majority and minority class samples equally. Using this algorithm, majority class samples were predicted accurately. So the authors introduced two methods i.e. (Data-Level Solution, Algorithmic level Solution) to overcome this situation. Data-level solution includes the sampling techniques i.e. Over-sampling and Under-sampling. Oversampling adds some samples to the minority class and under-sampling eliminates the samples of major class to make the data balanced. It was observed that there was drastic improvement in performance of area under receiver operating characteristics (AUROC) when ensemble or sampling techniques are used. For WEKA unbalanced dataset, the AUROC was proved from 0.432 to 0.982. It was also observed that though under-sampling reduced execution time but accuracy of classifying was distracted.

A novel approach of ensemble algorithm is proposed by Zhang Yongqing, et.al [4] with combination of hybrid sampling technique and committee of classifiers. The proposed approach can mitigate the problem of imbalanced data by rebalancing the data. The method proposed by authors can be applied on other fields in bioinformatics and perform well in protein-protein interaction but it could handle only gene micro array with more extra research.

Dr. Latesh Malik [5] discussed effective measures for improving the classification accuracy of skewed data streams. Data sampling method improves the classification accuracy of minority class but, because of infinite data streams and continuous concept drifting, this method cannot suitable for skewed data stream classification. Most existing imbalance learning techniques are solely designed for two class problem. Multiclass imbalance problem mostly solve by using class decomposition. AdaBoost with Negative Correlation (AdaBoost.NC) is an ensemble learning algorithm that combines the strength of negative correlation learning and boosting method. This algorithm is principally employed in multiclass imbalance data set. The results suggest that AdaBoost.NC combined with random oversampling can improve the prediction accuracy on the minority class without losing the overall performance compared to other existing class imbalance learning methods. However, this method improves the classification accuracy of minority class but, because of infinite data streams and continuous concept drifting, this method cannot suitable for skewed data stream classification.

A novel approach for learning from imbalanced datasets through combination of boosting and SMOTE introduced by Saumil Hukerikar, et.al [6]. The algorithm was illustrated by means of twenty real datasets and two of synthetically generated datasets having various features, percentage of imbalance and size. The datasets were chosen such that they sufficiently model the real- world scenario. The results obtained indicate that Skew Boost performs well against imbalanced datasets. In particular, algorithm proposed by authors has achieved comparable and slightly better performance in the measures like F-Measure, G- Mean and Area under Curve. On the basis of the results of real and synthetic datasets, the algorithm has applications in a wide range of fields, with medical and health care being the most prominent one. The algorithm gives consider- ably good results in other datasets from the science and technology domain also. The algorithm can serve well in real world applications like network intrusion detection, oil spill and identifying the very small number of patients having a rare disease. However, in cases of some datasets, the algorithm stays behind of some existing algorithms in metrics related to the majority class accuracy.

Many methods for alleviating the problem of class imbalance, including data sampling and boosting, which are the two techniques investigated by Chris Seiffert et.al [7]. The study addresses the issue of class imbalance, including an investigation of the types of imbalance that most negatively impact classification performance, and a small case study comparing several techniques for alleviating the problem. Data sampling has received lot of attention in analysis related to class imbalance. Data sampling attempts to overcome imbalanced class distributions by adding examples to (oversampling) or removing examples from the data set. The simplest form of undersampling is Random Under-Sampling (RUS); RUS arbitrarily removes examples from the majority class until a desired class distribution is found. While there’s no universally accepted best class distribution, a balanced (50:50) distribution is often thought to be near optimal. However, the complexity of algorithm gets increased with increased training time of model.

Son Lam Phung, et.al [8] reviewed existing approach for facing the problem of class imbalance and discussed various metrics to evaluate classifiers performance. They have proposed a new approach by reviewing existing approaches to deal with problem of class imbalance by combining both supervised and unsupervised learning. The proposed approach is a combination of supervised and unsupervised learning to handle imbalanced dataset which could be applied to existing training algorithms. The experiment done by authors on proposed approach can improve classification accuracy of minority class and classification performance effectively. However, the proposed approach has given higher values of G-means and F-measure than its counterpart while comparing the results of different training algorithms over all data-sets like Liver, Hepatitis, Pima Diabetes, Wisconsin and Breast Cancer.

A novel strategy for Support Vector Machine (SVM) in class imbalanced scenario was introduced by Kai Ming Ting, et.al [9]. They particularly, focuses on orienting the trained decision boundary of SVM so that a good margin between the decision boundary and each of the classes is maintained, and also classification performance is improved for imbalanced data. In contrast to existing strategies that introduce additional parameters, the values of which are determined through empirical search involving multiple SVM training, this strategy corrects the skew of the learned SVM model automatically irrespective of the choice of learning parameters without multiple SVM training. It compares the strategy with SVM and SMOTE, a widely accepted strategy for imbalanced data, applied to SVM on five well known imbalanced datasets. Also, it demonstrated improved classification performance for imbalanced data and is less sensitive to the selection of SVM learning parameters. However, it is not much clear as to how a particular value of these parameters affect the SVM hyperplane and the generalization capability of the learned model.

Nathalie Japkowicz, et.al [10] proposed a feature selection framework, which selects features for positive and negative classes separately and then explicitly combines them. It shows simple ways of converting existing measures so that they separately consider features for negative and positive classes. It uses a multi- strategy classifier system to construct multiple learners, each doing its own feature selection based on genetic algorithm. The proposed system also combines the predictions of each learner using genetic algorithms. It makes use of cluster-based oversampling to counter the effect of class imbalance and small disjuncts. However, feature selection can often be too expensive to apply.

Herna L. Viktor et.al [11] discussed a novel approach for learning from imbalanced data sets, DataBoost-IM, that combines data generation and boosting procedures to improve the predictive accuracies of both the majority and minority classes, without forswearing one of the two classes. The aim of the approach is to ensure that the resultant predictive accuracies of both classes are high. This approach differs from previous work in the following ways. Firstly, it separately identifies hard examples from, and generates synthetic examples for, the minority as well as the majority classes. Secondly, it generates synthetic examples with bias information towards the hard examples on which the next component classifier in the boosting procedures needs to focus. It provides additional knowledge for the majority as well as the minority classes and thus prevent boosting overemphasizing the hard examples. Thirdly, the class frequencies within the new training set are rebalanced to alleviate the learning algorithm’s bias toward the majority class. Rebalancing thus involves the utilization of a reduced number of examples from the majority and minority classes to ensure that both classes are represented during training. However, it has complex execution with tedious approaches.

Nitesh V. Chawla, et.al [12] suggested the design and implementation of an approach for the construction of classifiers from imbalanced datasets. They suggested that the combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) than only under-sampling the majority class. The method of over-sampling the minority class involves creating synthetic minority class examples. Experiments were performed using C4.5, Ripper and a Naive Bayes classifier. The proposed method was evaluated using the Area under the curve (AUC) and the Receiver Operating Characteristic (ROC) curve convex hull strategy. It was observed that a minority class sample could possibly have a majority class sample as its nearest neighbor rather than a minority class sample. However, this crowding lead to the redrawing of the decision surfaces in favor of the minority class.

Table I summarizes the highlights and observations of related work discussed above.

TABLE I

LITERATURE REVIEW

Ref. No.	Highlights	Observations
[1]	ARCID performs slightly better than Fitcare and RIPPER. However, it outperforms all standard rule-based and non-rule based approaches by many ranks.	Managing the overwhelming number of CARs generated from real- life datasets and Removing redundant rules conveys the same information.
[2]	1. EDBC uses the three methods to classify the imbalance data in proper in manner. 2. SplitBal works parallelly so the processing time is less as compare to EDBC.	1. EDBC does not works parallelly so the processing time is more. 2. In SplitBal dataset is divided into bins so working on bin is lengthy.
[3]	**1.** The combination of SMOTE + BAG + RF effectively improves the performance of classifying of imbalanced dataset. **2.** The use of ensemble and sampling techniques gives drastic improvement in performance in terms of AUROC	**1.** Execution time was reduced by using the under- sampling, but the accuracy of classifying dataset was distracted. **2.** There was loss of necessary and useful information due to undersampling.
[4]	1. SMOTE Method generates synthetic minority examples to over-sample the minority c1ass. 2. Bagging Method improves the performance of the overall system. 3. Support Vector Machine one of the most effective machine learning algorithms for many complex binary classification problems. **4.** SMOTE Bagging Method overcomes over-fitting problems of traditional algorithm.	1. SMOTE Method it only work on minority class not on majority class. Bagging Method combing of decision sometimes over- head.

[5]	Hybrid approach provides better solution for class imbalance. Suggests several algorithm and techniques that solve the issue of imbalance distribution of sample	This method improves the classification accuracy of minority class but, because of infinite data streams and continuous concept drifting, this method cannot suitable for skewed data stream classification.
[6]	SkewBoost performs better than DataBoost-IM, Inverse Random UnderSampling (IRUS), EasyEnsemble and Balance Cascade and Cost Sensitive Boosting (CSB2).	In case of some datasets, SkewBoost remains behind of some existing algorithms in metrics related to the majority class accuracy
[7]	RUSBoost presents an easier, faster, and fewer advanced alternative to SMOTEBoost for learning from imbalanced data.	The complexity and model training time get increased.
[8]	The approach proposed by author can effectively improve classification accuracy of minority classes while maintaining the overall classification performance.	The combination of supervised and unsupervised algorithm tends to have higher values of G-means and F-measure when compared with different training algorithms.
[9]	Z-SVM aims at finding the discriminating hyperplane that maintains an optimal margin from boundary examples called support vectors.	It is not much clear that how particular value of parameter affects the SVM hyperplane and generalization capability of learned model.
[10]	Use of cluster based oversampling counters the effect of class imbalance and small disjuncts.	Feature selection can often be too expensive to apply for handling imbalanced data.
[11]	DataBoost-IM produce s a series of high-quality classifiers will be better able to predict examples for which the previous classifier’s performance is poor.	The execution gets complicated when tedious approach is used.
[12]	For almost all the ROC curves, the SMOTE approach dominates. Adhering to the definition of ROC convex hull, most of the doubtless optimal classifiers are the ones generated with SMOTE.	A minority class sample could possibly have a majority class sample as its nearest neighbor rather than a minority class sample. This crowding will likely contribute to the redrawing of the decision surfaces in favor of the minority

		class.

Based on observations the proposed problem statement is to classify imbalanced data set to predict the accuracy of minority class from a given structured data set using diversified ensemble and clustering methods.

The main objectives of this work are:

- To develop balanced data set from imbalanced multi class dataset by applying under sampling techniques. To develop an efficient method for solving the class imbalance problem in the classification process.
- To increase accuracy rate of all classes at same time.
- To test and validate the model using existing approaches.

III. PROPOSED SYSTEM

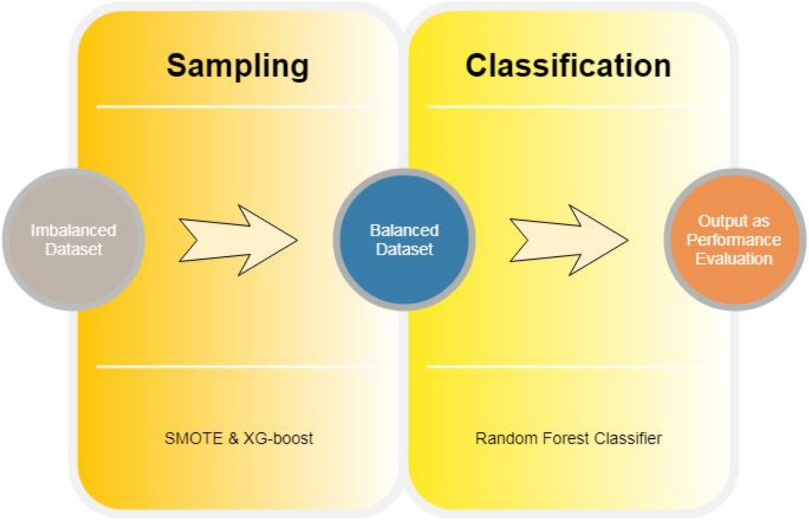


Fig. 2 System Design

In [3] the study about several solutions to handle imbalanced classification was given and various experiments for different techniques were performed. This paper considers two methods studied in [3] which are potentially suited for improving classification performance. They are XGBoost and SMOTE introduced in [1, 12] for handling imbalanced Dataset. One of the issues in classification is imbalanced dataset i.e. biased dataset. In this paper an application interface for handling the issue of imbalanced dataset is proposed. So that analyst doesn’t have to generate the model for handling imbalance between data from scratch.

The proposed work combines SMOTE, XGBoost, Random Forest Classifier. The main motive behind using SMOTE + XGBoost + Random Forest Classifier is not to lose useful and necessary data as in Random oversampling and split-up data up to maximum depth as provided and prune tree backwards reliably while providing more stable and accurate prediction. Main idea behind SMOTE + XGBoost is to generate synthetic minority instances by using KNN algorithm for each minority class instance [12] as illustrated in Fig. 3, rather than replicating the minority instances which causes overfitting as illustrated in Fig. 2. Using Random Forest Classifier this system will generate more stable and accurate prediction.

IV. RESULT & DISCUSSION

For Classification of imbalanced dataset, the following results were obtained –

A. User Interface

It contains list of menu items which can be accessed to have complete view of system. As this system is Machine Learning Model for handling Unbalanced dataset, the dataset is provided as input directly in the source code. The UI shows buttons for running Algorithm associated with them. The ‘Master’ button is used to load the dataset into the system. The ‘Random Forest’ button runs the Random Forest Algorithm along with hyper parameter tuning. The ‘XG-Boost’ button runs XG-Boost algorithm and shows it its optimizing. The ‘Data Manipulation’ button runs Multi Variate GMM module along with its adaptation to NG’s code from MATLAB. The ‘SMOTE’ button runs the proposed Machine Learning model namely ‘SMOTE+XG-Boost algorithm’ first then ‘RFC algorithm’ jointly to it. After running of each algorithm, the performance evaluation score in terms of Recall, Precision, and F1 score is viewed in output console of IDE. The Fig 3 shows the Home page of the system.

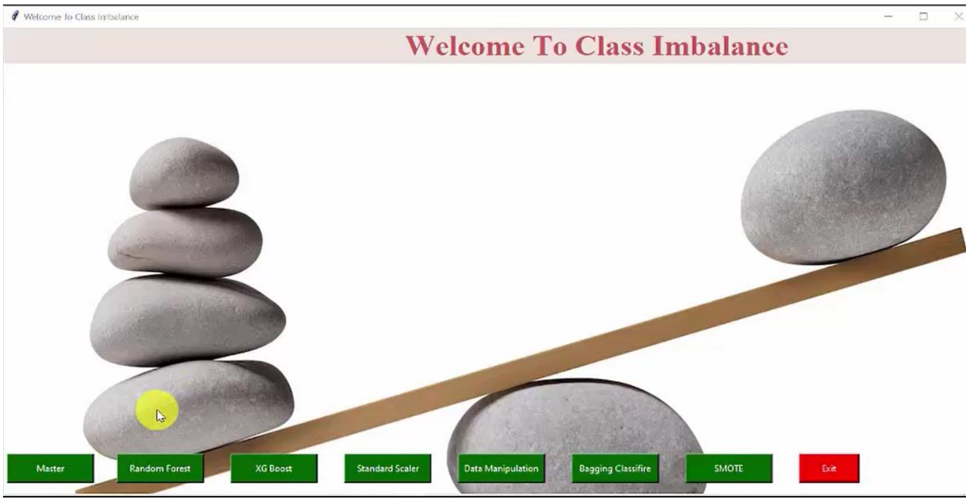


Fig. 3 User Interface

B. Distribution of Features in Dataset

The Fig. 4 illustrates the distribution plot of each features i.e. correlation of each feature in dataset.

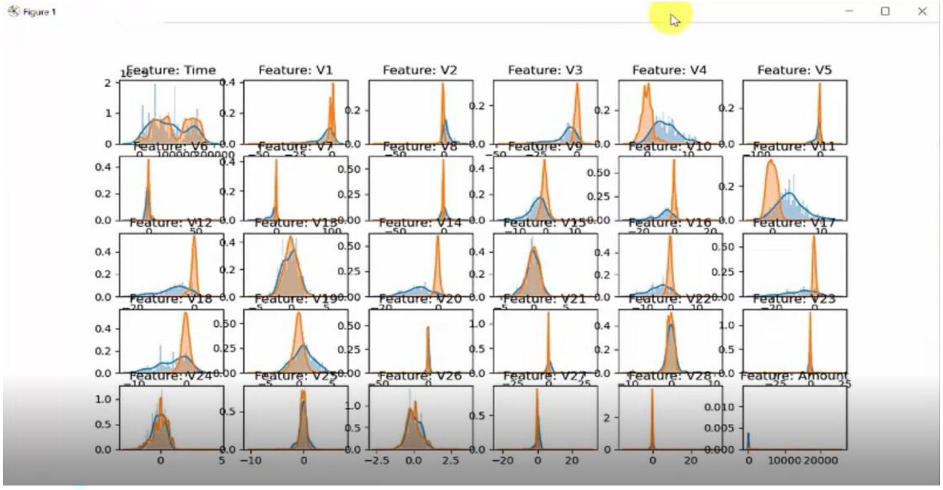
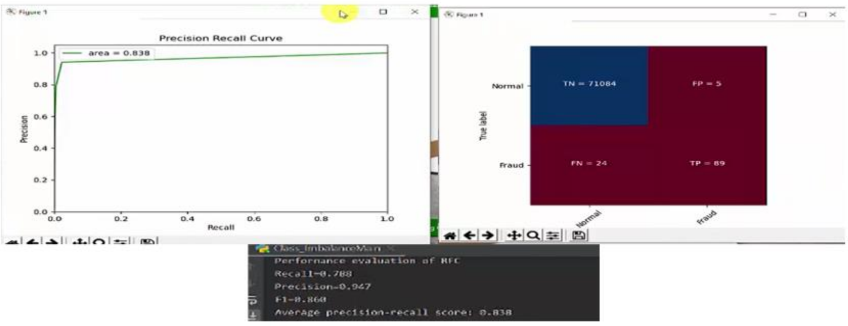


Fig. 4 Distribution plots of each features

C. Performance Evaluation of Algorithms

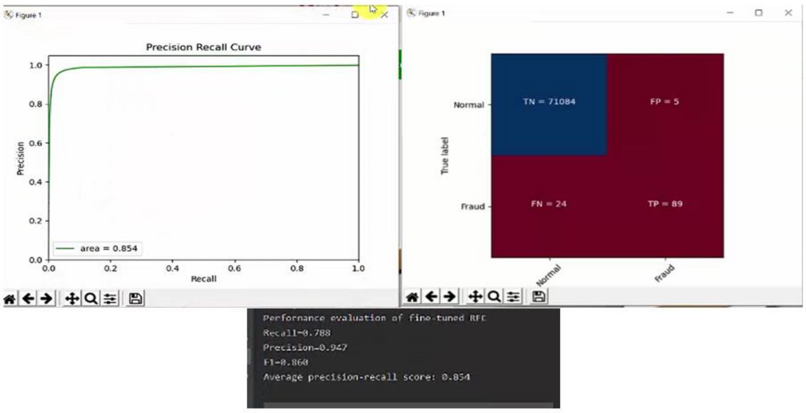
The Fig. 5 shows the performance evaluation scores of RFC along with their evaluation scores in terms of Recall value, Confusion Matrix, Precision Recall Curve, Precision value and F1 score.



```
Performance evaluation of RFC
Recall=0.788
Precision=0.947
F1=0.860
Average precision-recall score: 0.838
```

Fig. 5 Performance Evaluation of RFC

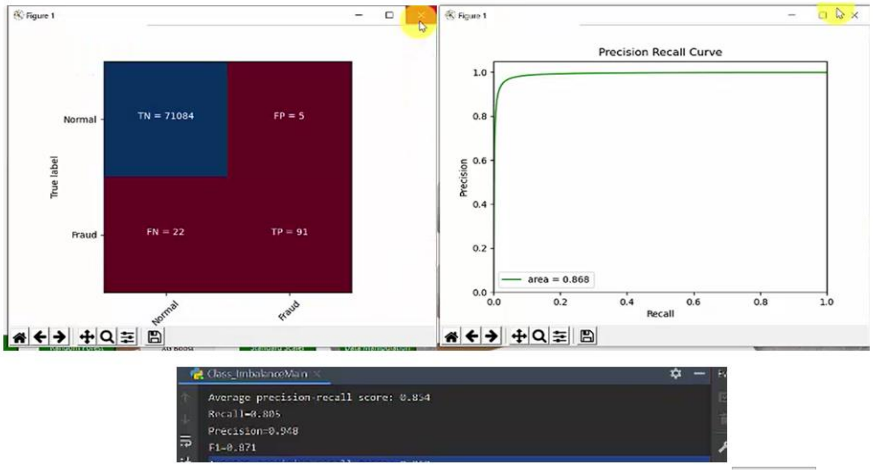
The Fig. 6 shows the performance evaluation scores of Tuned RFC and their evaluation scores in terms of Recall value, Confusion Matrix, Precision Recall Curve, Precision value and F1 score.



```
Performance evaluation of fine-tuned RFC
Recall=0.788
Precision=0.947
F1=0.860
Average precision-recall score: 0.854
```

Fig. 6 Performance Evaluation of Tuned-RFC

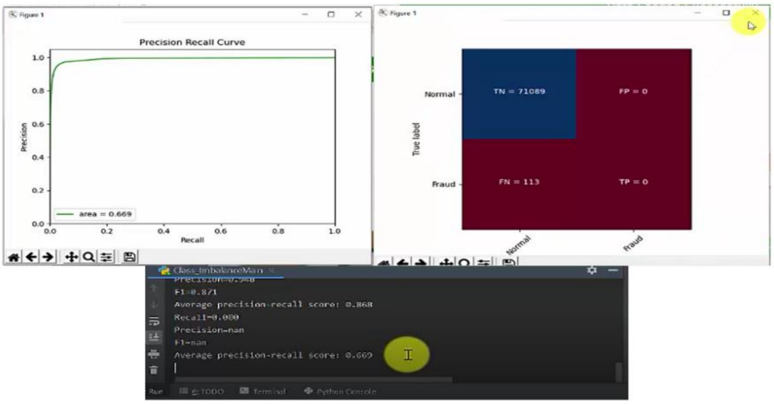
The Fig. 7 shows the performance evaluation scores of XG-Boost and their evaluation scores in terms of Recall value, Confusion Matrix, Precision Recall Curve, Precision value and F1 score.



```
Average precision-recall score: 0.854
Recall=0.805
Precision=0.948
F1=0.871
```

Fig. 7 Performance Evaluation of XG-Boost

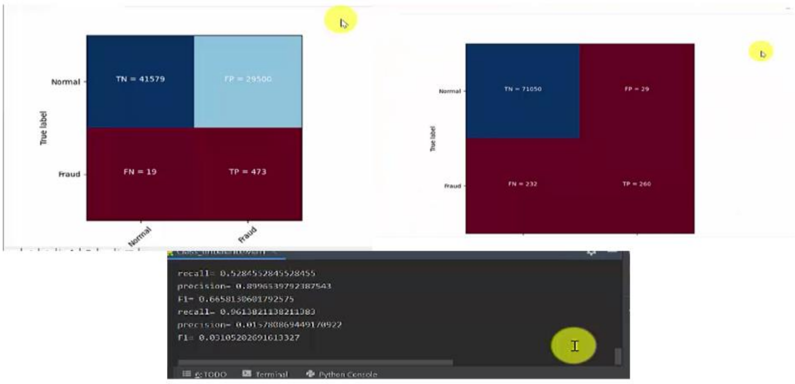
The Fig. 8 shows the performance evaluation scores of Optimized-XG-Boost and their evaluation scores in terms of Recall value, Confusion Matrix, Precision Recall Curve, Precision value and F1 score.



```
Precision=0.948
F1=0.871
Average precision-recall score: 0.868
Recall=0.000
Precision=nan
F1=nan
Average precision-recall score: 0.669
```

Fig. 8 Optimized XG-Boost

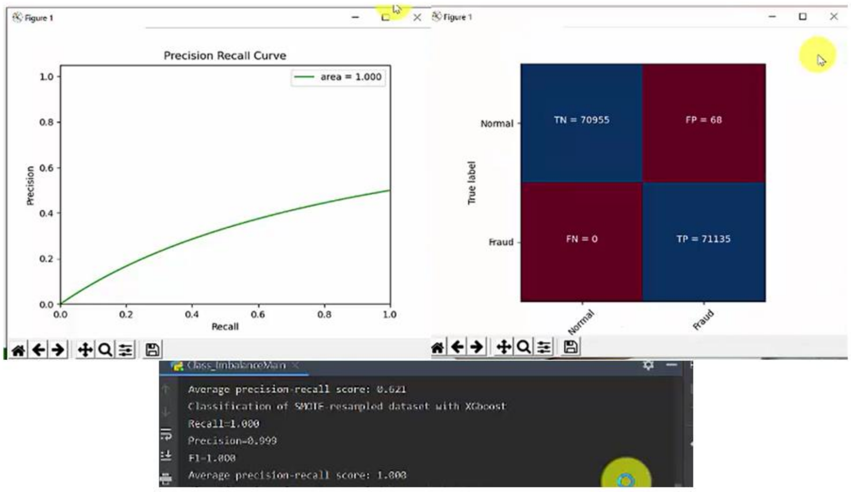
The Fig. 9 shows the performance evaluation scores of Multi Variate GMM and their evaluation scores in terms of Recall value, Confusion Matrix, Precision value and F1 score.



```
recall= 0.5284552845528455
precision= 0.8996539792387543
F1= 0.6658130601792575
recall= 0.9613821138211383
precision= 0.01578086944917092
F1= 0.03105202691613327
```

Fig. 9 Performance Evaluation of Multi Variate GMM

The Fig. 10 shows the performance evaluation scores of SMOTE+XGB and their evaluation scores in terms of Recall value, Confusion Matrix, Precision Recall Curve, Precision value and F1 score.

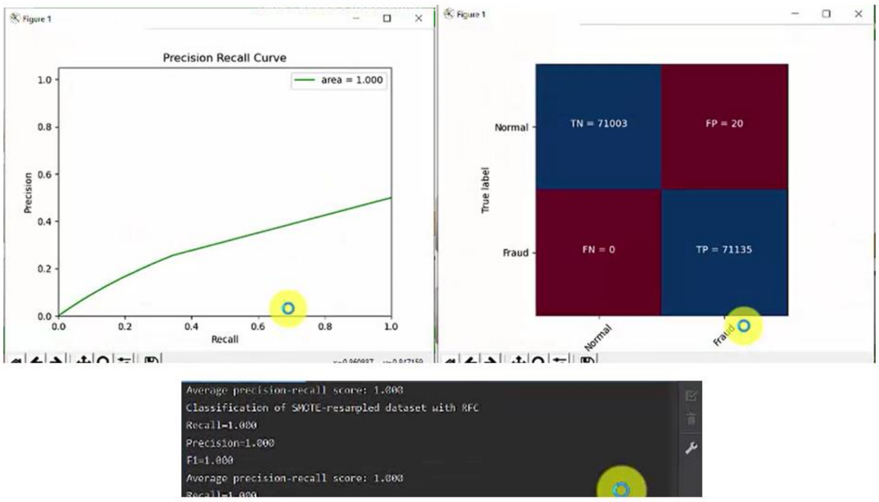


	Normal	Fraud
Normal	TN = 70955	FP = 68
Fraud	FN = 0	TP = 71135

```
Average precision-recall score: 0.621
Classification of SMOTE-resampled dataset with XGboost
Recall=1.000
Precision=0.999
F1=1.000
Average precision-recall score: 1.000
```

Fig. 10 Performance Evaluation of SMOTE+XGB

The Fig. 10 shows the performance evaluation scores of SMOTE+XGB+RFC and their evaluation scores in terms of Recall value, Confusion Matrix, Precision Recall Curve, Precision value and F1 score.



	Normal	Fraud
Normal	TN = 71003	FP = 20
Fraud	FN = 0	TP = 71135

```
Average precision-recall score: 1.000
Classification of SMOTE-resampled dataset with RFC
Recall=1.000
Precision=1.000
F1=1.000
Average precision-recall score: 1.000
Recall=1.000
```

Fig. 11 Performance Evaluation of SMOTE+XGBOOST+RFC

D. Final Result

By observing the performance evaluation of all the algorithms, we can see that SMOTE+XGBoost+RFC handles the imbalanced data better than RFC, XG-Boost & Bagging Classifier.

V. Conclusion

Classification is one among the foremost common problem in machine learning. One of the common issues found in dataset while classification is imbalanced classes issue. Various researchers have put their best efforts to solve this issue. The literature review has yielded some of the useful approaches to face this problem. The best solution to face this problem is to make use of ensemble methods. Ensemble methods are combination of multiple machine learning technique into one model to decrease variance (bagging), bias (boosting), or improve predictions (stacking). Other methods also exist but are not capable of performing best in classification of imbalanced dataset. In future, the proposed work will be implemented using python programming which will ensure that the target performance will be more than 95% so that ensemble classifier can be used reliably.

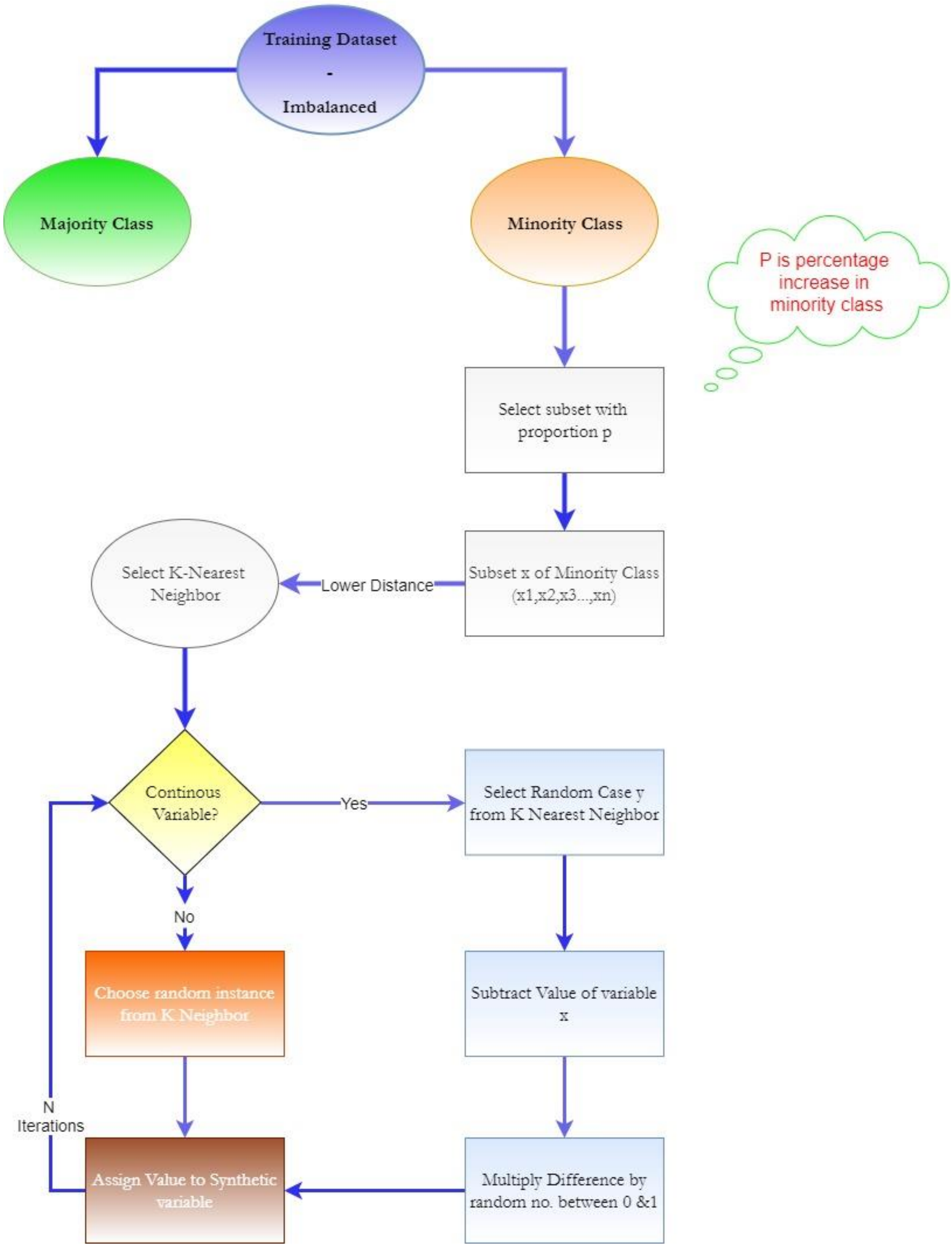

Training Dataset
-
Imbalanced
Majority Class
Minority Class
P is percentage increase in minority class
Select subset with proportion p
Subset x of Minority Class (x1,x2,x3...,xn)
Lower Distance
Select K-Nearest Neighbor
Continous Variable?
Yes
Select Random Case y from K Nearest Neighbor
No
Choose random instance from K Neighbor
Subtract Value of variable x
N Iterations
Assign Value to Synthetic variable
Multiply Difference by random no. between 0 &1

Fig. 12 Working of SMOTE

REFERENCES

[1] Abdellatif, Safa & Ben Hassine, Mohamed Ali & Ben Yahia, Sadok & Bouzeghoub, Amel. (2018). “ARCID: A New Approach to Deal with Imbalanced Datasets Classification.”, *SOFSEM 2018: Theory and Practice of Computer Science*, pp.569-580 doi:10.1007/978-3-319-73117-9_40.

[2] Jedrzejowicz, Joanna & Neumann, Jakub & Synowczyk, Piotr & Zakrzewska, Magdalena. (2017). “ Applying Map-Reduce to imbalanced data classification”. *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA*) pp.29-33. doi-10.1109/INISTA.2017.8001127.

[3] T. Lakshmi and S. Prasad, “A study on classifying imbalanced datasets,” *IEEE Conference Publication*, 2014.

[4] Yongqing, Z. & Min, & Zhu, Danling, & Zhang, M. & Gang, and M. Daichuan, “Improved SMOTEBagging and its application in imbalanced data classification,” *IEEE Conference Anthology*, 2013.

[5] M. R. Longadge, M. S. S. Dongre, D. L. Malik *et al.*, “Class Imbalance Problem in Data Mining: Review,” *International Journal of Computer Science and Network (IJCSN)*, vol. 2, no. 1, 2013. [Online]. Available: www.ijcsn.org

[6] S. Hukerikar, A. Tumma, A. Nikam, and V. Attar, “Skew Boost: An algorithm for classifying imbalanced datasets,” in *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, 2011, pp. 46–52.

[7] Seiffert, Chris & Khoshgoftaar, Taghi & Van Hulse, Jason & Napolitano, Amri. (2010). “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. Systems, Man and Cybernetics, Part A: Systems and Humans,” *IEEE Transactions*, 2010.

[8] S. L. Phung, “Learning pattern classification tasks with imbalanced datasets,” in *Pattern Recognition*. InTech, 2009.

[9] Imam, Tasadduq & Ting, Kai & Kamruzzaman, Joarder. “Z-SVM: An SVM for improved classification of imbalanced data,” *Advances in Artificial Intelligence*, vol. 4304, 2006.

[10] Chawla, Nitesh & Japkowicz, Nathalie & Kołcz, Aleksander. (2004). “Editorial: Special Issue on Learning from Imbalanced Data Sets.”, *SIGKDD Explorations*. 6. 1-6. 10.1145/1007730.1007733..

[11] Guo, Hongyu & Viktor, Herna. (2004). “Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach.” *SIGKDD Explorations*. 6. 30-39. 10.1145/1007730.1007736.

[12] N. Chawla, “Synthetic Minority Oversampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.