Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA
© Creative Research Publishers
Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

A Retrospect on Techniques for Solving Classification Problem of Imbalance Credit Card Dataset

Adepegba, OA.[1], Akinboro, S.[2], Olabiyisi, S.[3], Longe, O.B.[4], Amao, F.[5] Adepegba, S[6], & Aroyehun, A[1]
[1]Department of Computer Science, Adeleke University, Ede, Osun State, Nigeria
[2]Department of Computer Science, University of Lagos, Nigeria
[3]Department of Computer Science, Ladoke Akintola University, Ogbomoso Nigeria
[4]Department of Computational Sciences & Informatics Academic City University College Accra Ghana
[5]Department of Mathematics Adeleke University, Ede, Osun State, Nigeria
[6]Department of Computer Science, University of Ibadan, Nigeria

ABSTRACT

Classification of data becomes difficult because of unbounded size and imbalance nature of data. Class imbalance problem become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The minority samples are those that rarely occur but very important. There are different methods available for classification of imbalance data set which is divided into three main categories, the algorithmic approach, data preprocessing approach and feature selection approach. Each of this technique has their own advantages and disadvantages, for example, Credit card fraud datasets are naturally imbalanced by having more legitimate transaction in comparison to the fraudulent transactions. Literature represents numerous studies that are aimed to balance the skewed datasets. There are two major techniques of resampling in balancing these sets i.e. under-sampling and oversampling. However both under-sampling and oversampling techniques suffer from their own set of problems that can seriously affect the performance of classifiers that have been inducted for credit card studies in the past. Thus to accelerate prediction of credit card fraud, it is very important to implement the strategy that could possibly provide better predictive performance. This paper presents a comprehensive overview of resampling techniques that has been in use to solve the highly skewed data distributions for the domain of credit card fraud detection and to propose a better resampling (combining both undersampling and oversampling) techniques and the use of super learner approach using stacking method to improving the predictive performance.

Keywords: Oversampling, Undersampling, Hybrid Sampling, Resampling Techniques, Data imbalance

Adepegba, O.A., Akinboro, S., Olabiyisi, S.O, Longe, O.B., Amao, F. Adepegba, S, & Aroyehun, A. (2022): A Retrospect on Techniques for Solving Classification Problem of Imbalance Credit Card Dataset . Journal of Advances in Mathematical & Computational Sciences. Vol. 9, No. 2. Pp 87-92
Available online at www.isteams.net/mathematics-computationaljournal.

1. INTRODUCTION

In many real time applications large amount of data is generated with skewed distribution. A data set said to be highly skewed if sample from one class is in higher number than other. (Sou *et al.,* 2012). In imbalance data set the class having more number of instances is called as major class while the one having relatively less number of instances are called as minor class (Awoyemi *et al.*, 2017). Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions (Riffi *et al.,*2020), managing risk and predicting failures of technical equipment.

In such situation most of the classifier are biased towards the major classes and hence show very poor classification rates on minor classes. It is also possible that classifier predicts everything as major class and ignores the minor class. various techniques have been proposed to solve the problems associated with class imbalance (Singh and Jain 2019) , which divided into three basic categories, the algorithmic approach, data-preprocessing and feature selection To increase the prediction rate of minority class, a lot of studies in literature have given numerous resampling techniques. However there are two basic techniques which are followed widely.

These include Over-Sampling and Under-Sampling or combination of both Over-Sampling and Under-Sampling: which is called Hybrid Sampling.

- Under-Sampling: it removes the majority samples to the desired level of balance.

- Over-Sampling: generate new minority samples to the desired level of balance.

- Hybrid Sampling: This implements both over-Sampling and under-Sampling techniques until the desired level of balance is reached. There is wide range of resampling techniques that are implemented for credit card frauds but among them, three mostly used resampling techniques selected. In this study, a retrospective review of resampling techniques used by previous researchers is done. This paper presents a comprehensive overview of methods for solving the problem of data imbalance that further solves classification challenges on a dataset by profiling a solution of using a hybridized approach using both undersampling and oversampling techniques and the adoption of Super Learner ensemble Technique also known as stacking method.

This proposed technique adopts Genetic Algorithm for modelling a balancer, then is applied to the area of credit card fraud prediction on a highly imbalanced dataset and is expected to detect fraudulent transaction by classifying the transaction as fraud or as a legitimate transaction, a K-fold cross validation process is employed by making sure that all the dataset participates in both training and testing process.

2. REVIEW OF RESAMPLING TECHNIQUES FOR BALANCING IMBALANCE DATASET IN SOLVING CLASSIFICATION PROBLEM

In recent years, machine learning has gained a lot of traction in areas including image analysis, natural language processing, and speech recognition. Applying machine learning approaches to construct reliable fraud detection algorithms is crucial in this context for limiting losses and supporting fraud investigators, (Suryanarayana *et al., 2018)*. Yet there still exists challenges with dataset imbalance and this can be by profiling solution (combining oversampling and undersampling) techniques to address the issue.

Table 1: Comparison Table of Previous Research works

S/N	Author	Year	Dataset	Methodology	Contribution	Limitation
1	Kulkarni and Ade	2016	German Credit Card dataset	Logistic Regression	Maintained efficiency	Model is complex
2 2	Fu, K et al	2016	Real-world Dataset	CNN	Recognize complicated fraud patterns	Dataset was highly imbalanced
3	Seera, et al	2017	European Cardholder	Majority Voting+ Adaboost	Obtaining best result using MCC metric	MV is unstable in the absence of noise
4	Alex, et al	2018	PagSeguro transactions data	Bayesian network classifier (BNC)	Increase the accuracy and economic efficiency	Results are obtained only in terms of F1 metric
6	Vardhani et al	2019	European Cardholder	Smote	Selects the best features and reduce the overfitting	High computational
5	Naoufal, and Nourddine	(2020)	Random Forrest Classifiers (RFC)	Selection Features and Support Vector Machine for Credit Card Risk Identification	The AUC-Roc measure also validates that the proposed method has the best performance and skills in identifying if the transaction presents a risk or not	Unbalanced dataset was a challenge
7	Fayaz- Itoo et al	2020	European Cardholder	Random Over Sampling	Improve the accuracy and classification time	The same data is missing in Random under Sampling

3. ENSEMBLE LEARNING TECHNIQUES OF SOLVING DATA IMBALANCE PROBLEM

A base learner is a combination of numerous learners that make up an ensemble. A collection of various strengths in each base learner, followed by work on their weaknesses, results in an ensemble that is more efficient, effective, and robust than each of the base learners, who are worthy problem solvers in their own right in each of their classes, resulting in providing a more reliable solution (Eweoya *et al.,* 2019).

Rasim et al (2020) adopted bagging classifier based on the decision tree works well with this kind of data since it is independent of attribute values the challenges in this work was; non-availability of real data set, Unbalanced data set, size of the data set, determining the appropriate evaluation parameters, dynamic behaviors of fraudsters.

In Enhancing the Credit Card Fraud Detection Through Ensemble Techniques used Decision tree, Naives Bayes, support vector machines, boosting and Adaboost, the Boosting and Decision Tree resulted with the highest accuracy of 98.37% and F -Measure of 94.49% while Decision Tree only resulted with accuracy of 98.27% and F -Measure of 93.98%, the limitation with this work was the use of accuracy and F- measure score as the only performance evaluation metrics. Different ensemble approaches result from the use of different base learner generation processes or different combination schemes. In the literature on ensemble, the three representative effective and foundational ensemble methods are boosting, bagging, and stacking. AdaBoost's weighting strategy is equal to resampling the data space, which may be used by most classification systems without requiring them to change their learning algorithms. Furthermore, it could minimize the additional learning costs associated with determining the best class distribution and representative samples. Furthermore, it decreases information loss and the risk of overfitting when compared to the process of removing samples from a data set.

4. SPECIFIC CHALLENGES IN PREVIOUS WORK

One major challenge is that when working with a large amount of data, it's possible to encounter data sparsity problem, which means that some data points are missing and this can adversely affect the system's efficiency.

5.CONCLUSION AND FUTURE WORK

This review investigated different resampling techniques used for addressing the issue of data imbalance. Some limitations were identified which include unavailability of large dataset, some data were highly imbalance, use of smaller dataset was also a challenge, highly negatively skewed data and so on (Olowookere and Adewale 2020). Thompson *et al.,* (2019) adopted SMOTE (Synthetic Minority Over -sampling Technique) for solving data imbalance problem, Ensemble methods and SMOTE technique was also used by (Sood, 2020) to oversample training data with positive classes, so that classifiers have more positive class to work with, but the limitation with SMOTE is while generating synthetic examples, SMOTE does not take into consideration that neighboring examples can be from other classes whereas, this can increase the overlapping of classes and can introduce additional noise. Olowookere and Adewale, (2020) adopted cost sensitive method to checkmate data imbalance the cost-sensitive learning used here treats the different misclassifications differently by invoking a cost-sensitive Logistic Regression algorithm on the training set. Meanwhile, the logistic regression algorithm here is in its standard form and does not take into consideration the varying misclassification costs (that is, it is cost-insensitive).

Therefore there is need for increase in classification performance accuracy and in other evaluation metrics to be in use in a particular subject area. Many researchers suggest combining oversampling and undersampling methods to balance the dataset better. Hence this work suggests modeling a balancer by adopting a hybridized approach of combining both over sampling and downsampling techniques using a Voter for Super Learner Ensemble technique to solve the problem of data imbalance in a labeled dataset such as that of the credit card data.

The ensemble-based method is another technique which is used to deal with imbalanced data sets, and the ensemble technique is combining the result or performance of several classifiers to improve the performance of single classifier. This method modifies the generalization ability of individual classifiers by assembling various classifiers. It mainly combines the outputs of multiple base learners and then using a voter based on the output of the multiple base learners.

The adoption of the Voter for Super learner Ensemble couple with the modeled enhanced Genetic Algorithm oversampling and downsampling balancer, this will help to prevent misclassification, eliminate noise in the imbalanced data sets by this, a significant research gap is filled with regards to the subject area.

REFERENCES

1. Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9. IEEE.
2. Barahim, A. Alasaibia, N. Alhajri, A. and Mohammed, N. (2019) Enhancing the Credit Card Fraud
3. Detection through Ensemble Techniques Article in Journal of Computational and Theoretical Nanoscience November 2019 DOI: 10.1166/jctn.2019.8619 (2019)
4. Eweoya, I., Adebiyi A., Azeta, A., & Olatunji O., (2019) “Fraud Prediction in Bank Credit Administration: A Systematic Literature Review” Journal of Theoretical and Applied Information Technology 15th June 2019. Vol.97. No 11 ongoing JATIT & LLS ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195 3147
5. Fu, K., Cheng, D., Tu, Y., and Zhang, L. (2016 October) Credit card fraud detection using convolutional neural networks. Springer, pp. 483-490.

6. Kulkarni, P., & Ade, R. (2016) Logistic regression learning model for handling concept drift with unbalanced data in credit card fraud detection system. In Proceedings of the Second International Conference on Computer and Communication Technologies, pp. 681-689. Springer, New Delhi.
7. Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., and Zeineddine, H. (2019) An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, pp. 93010-93022.
8. Naoufal N., Nourddine E., (2020) '*Selection Features and Support Vector Machine for Credit Card Risk Identification*' Volume 46, 2020, Pages 941-948 https://doi.org/10.1016/j.promfg.2020.05.012
9. Olowookere T., & Adewale O., (2020) '*A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach*'. Scientific African Journal Volume 8 (2020) e00464 https://doi.org/10.1016/j.sciaf.2020.e00464 2468-2276
10. Rashmi, S. Chetan, J. Suresh, k. (2020) 'Credit Card Fraud Detection Using Supervised Learning Approach', International Journal of Scientific & Technology Research Volume 9, Issue 10, October 2020 ISSN 2277-8616 216 IJSTR (2020)
11. Riffi, J., Mahraz, M. A., El Yahyaouy, A., & Tairi, H. (2020, June) Credit Card Fraud Detection Based on Multilayer Perceptron and Extreme Learning Machine Architectures. In 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1-5. IEEE.
12. Seera, K., Lim, M., and Nandi, A. (2018) Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, pp. 14277-14284.
13. Singh, A., & Jain, A. (2019) Adaptive credit card fraud detection techniques based on feature selection method. In Advances in computer communication and computational sciences, pp. 167-178). Springer, Singapore.
14. Sood A., (2020) '*Credit Card Detector, Using Ensemble Methods and SMOTE Samping*,' 2017. https://www.kaggle.com/ganakar/using-ensemble-methods-and smote-sampling. [Online]. Available: Accessed 20 April 2020
15. Thompson, A., Aborisade, L., Oyinloye, O., & Odeniyi, E. (2019) '*A Fraud Detection Framework using Machine Learning Approach* ' *IARA,* 2019 ISBN:978-1-61208-743-6
16. Taha, A. A., & Malebary, S. J. (2020) An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, pp. 25579-25587. 3. Randhawa, K., Loo, C.
17. Vardhani, P., Priyadarshini, Y., and Narasimhulu, Y. (2019) CNN data mining algorithm for detecting credit card fraud. In Soft Computing and Medical Bioinformatics, pp. 85-9. Springer, Singapore.