International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I3P105

Volume 68 Issue 3, 21-27, March 2020

Original Article

Credit Card Fraud Detection from Imbalanced Dataset Using Machine Learning Algorithm

Swati Warghade[1], Shubhada Desai[2], Vijay Patil3

[1,2,3] *Information Technology Department, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India.*

Received Date: 31 January 2020
Revised Date: 15 March 2020
Accepted Date: 16 March 2020

Abstract *– In Today's world, a credit card is the most accepted payment mode for both online as well as offline. It provides cashless shopping at every shopping mall. It is the most convenient way to do online transactions. Therefore, the risk of fraud in credit card transactions has also been increasing.*

With the growing usage of credit card transactions, financial fraud crimes have also been drastically increased, leading to the loss of huge amounts in the finance industry. Having an efficient fraud detection algorithm has become a necessity for all banks in order to minimize such losses. In fact, the credit card fraud detection system involves a major challenge: the credit card fraud data sets are highly imbalanced since the number of fraudulent transactions is much smaller than the legitimate ones. This paper aims at analysing various machine learning techniques using various metrics for judging various classifiers. This model aims at improving fraud detection rather than misclassifying a genuine transaction as fraud.

Keywords *- Credit Card Fraud Detection, Imbalanced dataset, SMOTE.*

I. INTRODUCTION

Credit Card fraud is a measure problem arising in the online sector. Credit card fraud is referred to fraud or theft of someone else's credit card as used for payment for goods or receiving funds in cash. Some cardholders do not know that they are victims of credit card fraud until massive damage has already been done to their credit. It can take years for some people to recover from the damage that comes with being a victim of

Credit card fraud. So it is most preferable to take prevention than cure. Most machine learning algorithms work best when the number of instances of each class is roughly equal.

When the number of instances of one class far exceeds the other, problems arise.

Fraud detection in credit cards is truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve credit card fraud detection such as genetic algorithm, Artificial neural network frequent, regression, decision tree and the random forest is carried out.

Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal features selection for the models is the most important part of data mining to evaluate the performance of techniques on skewed credit card fraud detection. The learning phase and the Prediction of machine learning algorithms can be affected by the problem of an imbalanced data set.

To reduce the variance in the dataset, ensemble methods can be used so that they give effective and accurate results for the same.

The Ensembling method is useful to overcome a difference in the population of instances differences in hypothesis.

A. Data Pre-processing

As the data working on is highly imbalanced and noisy. We have to preprocess the data so that our model can be efficiently trained. There are many approaches to handle this problem of Imbalanced Datasets, such as Oversampling of ensemble method. Since the dataset we have used contains several features, it might have a lot of variances. So data has also been transformed so that it has nearly 0 variances. This leads to a better generalization of the model while training.

II. LITERATURE REVIEW

Paper No: 1 Study Report Of Using Genetic Algorithm in improve Classification of imbalanced Datasets for credit card fraud detection. (2018)

Method Used: Genetic Programming, K-means algorithm.
TOOL: Java IDE
Technology: Machine Learning
Description:

- This paper aims first: to enhance the classified performance of the minority of credit card fraud instances in the imbalanced data set. This paper proposes a sampling method based on the K-means clustering and the genetic algorithm.

- In the proposed paper, the k-means algorithm is used to split the minority class of fraud instances into clusters according to their similarities and generate new samples in these clusters.
- K-means clustering and genetic algorithm is an ensemble strategy.
- Genetic algorithms have been applied to handle imbalanced datasets by generating new minority class instances.
- Applying this algorithm to the bank credit card fraud detection dataset aims to reduce fraudulent transactions and decrease the number of false alerts.
- Further work is to implement this approach using a python programming language.

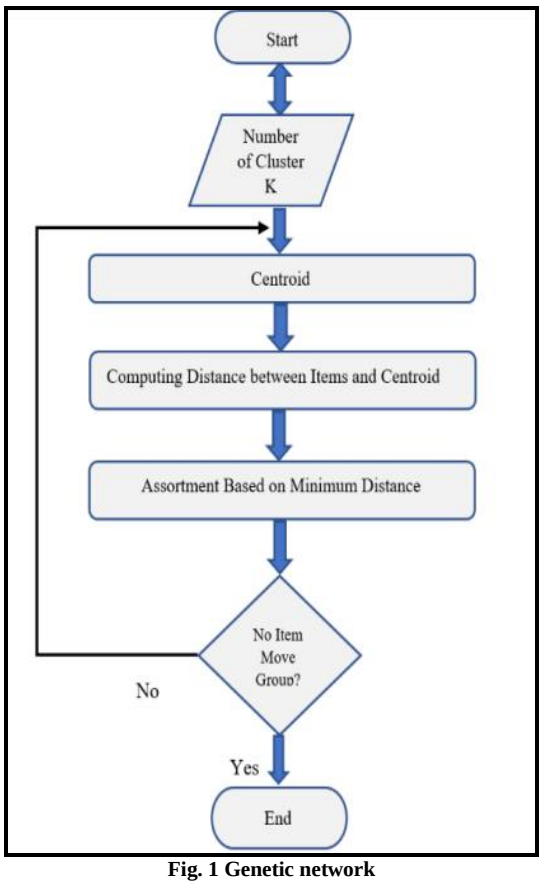

Fig. 1 Genetic network

Paper No: 2 Using deep networks for fraud detection in credit card detection.

Method: Auto-encoder, Deep networks.
Technology: Deep learning
Description:

One of the most interesting subjects that need more attention in prediction accuracy is fraud detection. As the deep network can gradually learn the concepts of any complicated problem, using this technic in this domain is very beneficial. We propose a deep auto-encoder to extract the best features from the information of the credit card transactions and then append a SoftMax network to determine the class labels.

Methods to detect fraudulent activities:-

- K-nearest neighbour (KNN) with association rules.
- Genetic Programming to evolve decision trees for data classification and prevent the search spaces from becoming extremely large.
- Self-organization map algorithm to map data in a discriminative space.
- We can pass the initial features of each transaction to the network and, by training a deep auto-encoder, extract the appropriate features.
- The structure of such a network depends on the number of available features in the dataset. After that, we can use a softmax layer to decide about the class label.

Paper no: 3 Study of Hidden Markov Model in Credit Card Fraudulent Detection

Model: HMM Model
Technology: Machine learning
Description:

- Hidden Markov Model will be helpful to find out the fraudulent transaction by using the spending profiles of the user. It works on the user spending profiles, which can be divided into major three types such as:-
- Lower profile; Middle profile; higher profile.
- For every credit card, the spending profile is different, so it can figure out an inconsistency of the user profile and try to find the fraudulent transaction. A Hidden Markov Model is a finite set of states. Transitions among these states are governed by a set of probabilities called transition probabilities. In a particular state, a possible outcome or which is associated symbol of the observation of probability distribution.

A. Comparison between Literature Surveys

Paper	Paper 1	Paper 2	Paper 3
Algorithm	K-Means Algorithm, Genetic Algorithm	Autoencoder	Hidden Markov Model
Technology	Machine learning	Deep Learning	Machine learning
Pros	It is efficient, easy and fast for small datasets.	Ability to face big data set	Helps to obtain high fraud coverage combined with false alarm rate

Cons	KNN can require a lot of memory or space to store all data	Not suitable as general-purpose algorithms because they require a very large amount of data	Large no of unstructured parameters. Cannot express dependencies between hidden states.

Fig. 2 Comparison of papers

III. PROBLEM STATEMENT

In the Digital World, everyone prefers online shopping online payment. Because it is efficient and convenient as well as easy for the customers. But crimes and frauds are always come with technology. Fraud and crime lead to easy earning of money. The risk associated with credit card fraud is related to measure issuing serious economic threats loss of personal information. Increasing the use of credit cards leads to fraud crimes, so Machine Learning gives an efficient way to detect fraud.

The standard algorithms are well performed towards the side of the majority class. So they predict only majority class data. Due to this minority class get ignored as algorithm treat it as noise. A large dataset contains majority instances as well as minority instances called imbalanced skewed data. Such data require additional precautions.

Therefore, methods to further improvement of credit card fraud detection and speed up are required to create a model which aims to improve fraud detection rather than misclassifying a genuine transaction as fraud.

IV. PROPOSED SYSTEM

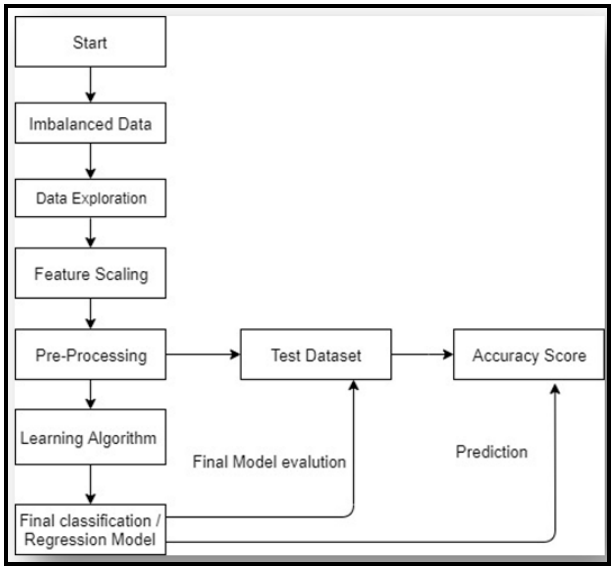

Fig. 3 The proposed system of credit card fraud detection

A. Data Imbalance and its effects on Prediction

An imbalance set is one where the number of instances belonging to one group is significantly higher than another group. Machine Learning algorithms tend to produce unsatisfactory classifiers when faced with imbalanced datasets. Due to imbalanced data, event prediction belongs to the majority class. Minority classes get ignored as noise due to their lower event rate compared to the majority class.

a) Oversampling

In this sampling technique, we randomly increase the minority class by randomly replicating the samples in the minority class, so the data gets well balanced and able to train by the various classification algorithms. But as we are increasing the number of instances of the minority class, it increases the time to train the model and the problem of overfitting.

Due to the overfitting problem, the additional data, if any comes to us, will be difficult to fit in the same curve, so the accuracy of the model will get reduced.

B. Synthetic Minority Oversampling Technique (SMOTE)

It uses both undersampling and oversampling to balance the existing dataset. But, unlike traditional oversampling Algorithms, it doesn't just replicate the minority class instances. It uses KNN (K Nearest Neighbours) for selecting the instances of the minority class. Then it randomly chooses Instances from the minority class create new instances of minority class by using the selected instance feature's vector and the difference between the N features vectors multiplied by a random number.

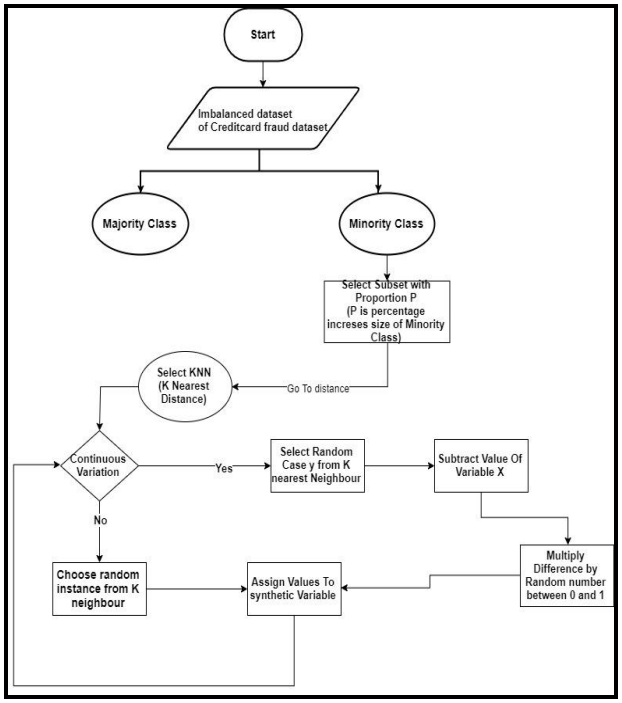

Fig. 4 SMOTE

This approach for oversampling has been quite popular and, in some cases, help in improving the model accuracy. But it has several disadvantages. Though it avoids the problem of overfitting, the data generated is synthetic and might not be in resemblance to the original data.

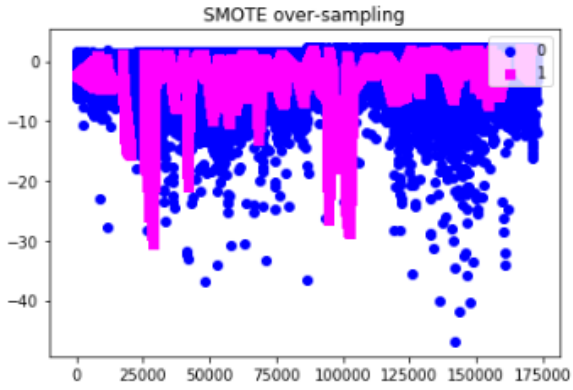

Fig. 5 Smote Oversampling

a) Undersampling

In this sampling technique, we randomly select the majority class to extract a smaller set of it and preserve it for the minority class. We add the number of majority instances in such a way that the majority and minority class ratio becomes 1:1 as our aim is to detect fraud transactions more precisely and also improve the accuracy of the model. So undersampling helps us to achieve better fraud detection as it doesn't make any changes in fraud data. Data undersampling made the data balanced by making the ratio of majority and minority classes 1:1. Some data from the legal transactions are mixed with the fraud ones.

V. PERFORMANCE ANALYSIS AND CLASSIFICATION TECHNIQUES

A. Local Outlier Factor

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method that computes the local density deviation of a given data point with respect to its neighbours. LOF is based on concepts of density based detection. It is a method to find similarities and dissimilarities between factors and variants, which helps to correct outlier detection. It ensembles LOF factors to give patterns to detect outliers in the environment.

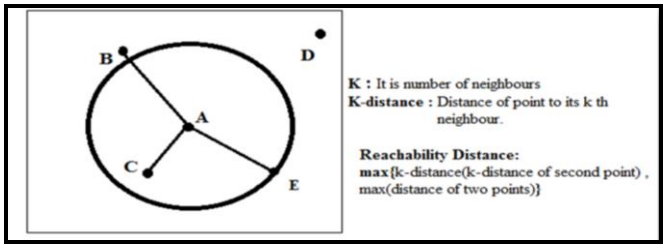

Fig. 6 Local Outlier Factor

- In the LOF algorithm, the difference in density between a data object and its neighbourhood is the degree of being an outlier, known as its local outlier factor. Intuitively, outliers are the data objects with high LOF values, whereas data objects with low. LOF values are likely to be normal with respect to their neighbourhood.

B. Isolation Forest

The Isolation Forest algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. To avoid issues due to the randomness of the tree algorithm, the process is repeated several times, and the average path length is calculated and normalized. Isolation Forest is an outlier detection technique that identifies anomalies instead of normal observations

C. Support Vector

SVM or Support Vector Machine is a linear model for classification and regression problems. According to the SVM algorithm, we find the points closest to the line from both classes. These points are called support vectors.

$$z = x^2+y^2$$

It works really well with a clear margin of separation. It is effective in high dimensional spaces. It is effective in cases where a number of dimensions are greater than the number of samples.

VI. EXPERIMENT

Accuracy Matrix and prediction model:

- **Precision** is a ratio of correctly predicted positive observation to total predicted positive observation
  Precision = TP/TP+FP
- **Recall** is the ratio of correctly predicted positive observations to all observations in actual class
  Recall = TP/TP+FN
- **F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
- **F1 Score = 2*(Recall * Precision) / (Recall + Precision)**
- **Support** is the number of occurrences of each class in y_true.
- **Micro Avg:** Calculate metrics globally by counting the total true positives, false negatives and false positives.
- **Macro Avg:** Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.
- **Weighted Avg:** Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance;

Comparative chart							
Algorithm		Precision	recall	F1 score	support	Accuracy Score	Fraud outlier
Local Outlier Factor	0	1	1	1	28432	0.9965	97
	1	0.02	0.02	0.02	49		
Support Vector Machine	0	1	0.46	0.63	28432	0.4584	15425
	1	0	0.46	0	49		
Isolation Forest	0	1	1	1	4987	0.9978	77
	1	0.22	0.22	0.22	49		

Fig. 7 Results

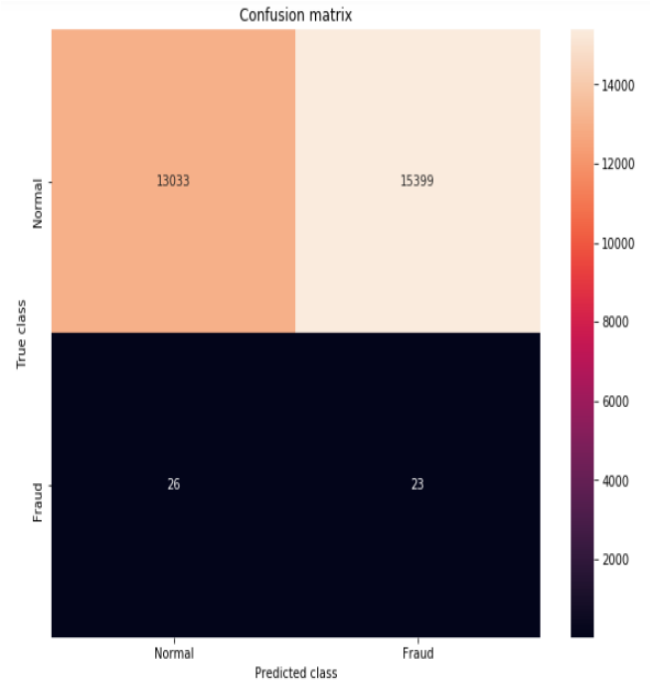

Fig. 8 Confusion Matrix

VII. CONCLUSION

In this model, synthetic techniques like SMOTE will perform the conventional oversampling method. For better results, one can use synthetic sampling methods like SMOTE along with advanced boosting methods like Local Outlier factor, Isolation Forest and SVM method.

Due to the parallel processing model, LOF and Isolation Forest is fast and robust to an outlier. By testing with samples of small records, change in percentage of training and testing dataset like 70/30, 80/20, 90/10 comparison of accuracy shown by following line graph. Null values affect the prediction scores of the model, so different techniques are used to remove the null values and then predict the accuracy of a model given in the below table. Isolation Forest gives a 99.74% accuracy score, and Support Vector Machine gives a 45.84% accuracy score, LOF gives a 99.66% accuracy score which makes the Prediction true rather than misclassifying the genuine transaction as fraud.

Accuracy Score					
Algorithm	70/30	80/20	90/10	Null Values	Sampling with small records
Support Vector Machine	45.84%	45.85%	45.84%	37.37%	48.92%
Isolation Forest	99.73%	99.73%	99.73%	99.64%	99.78%
Local Outlier Factor	99.66%	99.66%	99.66%	99.26%	99.46%

Fig. 9 Accuracy Scores

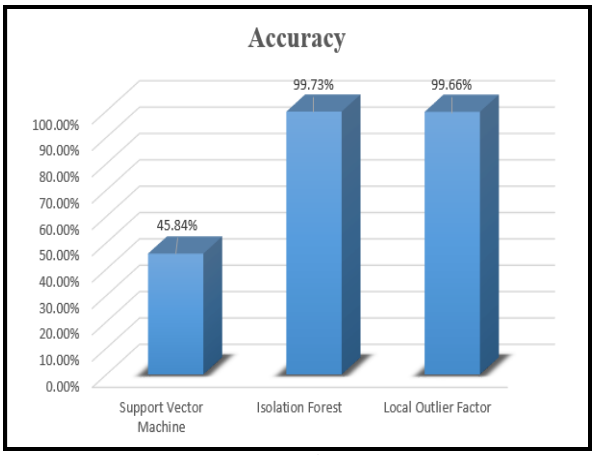

Fig. 10 Bar graph of accuracy score

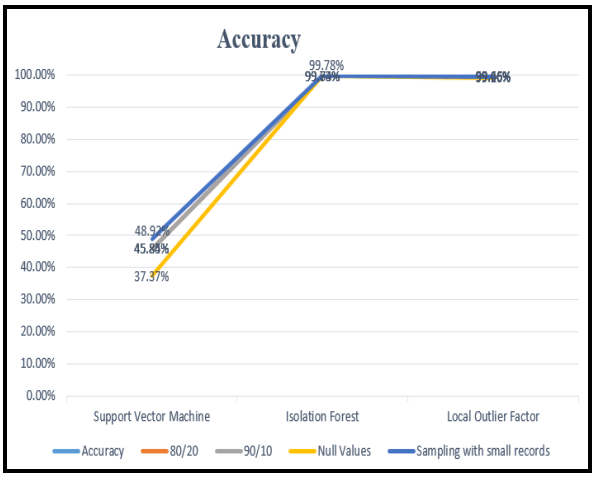

Fig. 11 Line graph of accuracy score

ACKNOWLEDGEMENT

The authors would like to thank the Guide and mentors for useful guidance as well as reviewers for helpful comments. Thanks to IJCTT for giving a platform to explore talents and useful style templates.

REFERENCE

[1] Ibtissam Benchaji, Samira Douzi , Bouabid El Ouahidi . Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection , 2nd Cyber Security in Networking Conference (CSNet)., (2018).

[2] Sahil Dhankhad, Emad Mohammed, Behrouz Far. Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: Comparative Study, IEEE., (2018).

[3] Ankit Mishra, Chaitanya Ghorpade., Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques., IEEE International Students' Conference on Electrical, Electronics and Computer Science ., (2018).

[4] Yasmirah Mandasari Saragih, Andysah Putera Utama Siahaan, Cyber Crime Prevention Strategy in Indonesia, SSRG International Journal of Humanities and Social Science., 3(6) (2016).

[5] Chunzhi Wang, Yichao Wang, Zhiwei Ye, Lingyu Yan, Wencheng Cai, Shang Pan., Credit card fraud detection based on whale algorithm optimizes HG BP neural network, The 13th International Conference on Computer Science & Education. Colombo, Sri Lanka, (2018)

[6] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang., Random Forests for Credit Card Fraud Detection., 978–1–5386–5053–0/18/$31.00© IEE., (2018).

[7] Haibing Li, Wing-Lun Lam, Chi-Wai Chung, Man-Leung Wong, Financial Fraud Detection: Multi-Objective Genetic Programming with Grammars and Statistical Selection Learning., SSRG International Journal of Computer Science and Engineering ., 7(2) (2020)

[8] https://www.geeksforgeeks.org/ml-credit-card-fraud detection/

[9] https://www.kaggle.com/bonovandoo/fraud-detection-with-smote-and-xgboost-in-

[10] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

[11] https://www.guru99.com/supervised-vs-unsupervised-learning.html