Filtering the Credit Card Fraud Detection Dataset for enhancing the Classification Performance

Neha Purohit and Dr. Rajeev G. Vishwakarma
Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam University,
Indore (M.P.) 452010, India
Corresponding Author Email : nehapurohit059@gmail.com

Article Info
Page Number: 10122-10130
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received:
12 September 2022
Revised: 16 October 2022
Accepted: 20 November 2022
Publication: 25 December 2022

Abstract

Due to flexibility of payments the utilization and acceptance of credit cards is growing day by day. Additionally, the government is also forcing to transect through the online channels. This will help to improve transparency in payment systems. But the cases of financial fraud are also increasing. There available systems for credit card fraud detection are suffering to deal with the highly noisy data. Therefore we proposed a dataset quality enhancement technique. The proposed technique utilizes chi-square test, missing value handling, MTD technique for dimensionality reduction and regression technique for outlier detection. After enhancing the quality of data we utilize the refined data with the two machine learning algorithms namely Xgboost and Convolutional Neural Network (CNN). The experiments are carried out and performance is measured in terms of accuracy and training time. The comparative results demonstrate the Xgboost and CNN both are providing accuracy up 99% but the time utilization of CNN model is higher as compared to XgBoost.

Keywords: Machine Learning, Classification, Credit Card Fraud Detection, Machine Learning Application, Supervised and Unsupervised Learning, Comparison.

I. INTRODUCTION

Credit card is one of the leading payment methods. The credit cards are accepted worldwide and can be utilized in different kinds of transactions. But sometimes the transactions are made through secure channels and sometimes the payment is done through the unsecured channels. The unsecured payment channels are highly risky to compromise with financial fraud through false credit card transactions. Such kind of fraud transactions can damage the reputation of banking company as well as causes the significant financial losses. Therefore, credit card detection is an essential application of banking system. In this context, a number of works is done by contributors to detect a fraud transaction. Most of the work is based on Machine Learning (ML). The ML techniques have the ability to analyse large amount of data and accurately detect the required information. But the source of experimental dataset is limited and dataset is noisy. Therefore, we need an effective technique to accurately detect the fraud transaction.

In this paper, the main aim is to study the data noise and outlier detection problems in credit card fraud detection. The outlier is a classical machine learning issue in data analysis. The sudden spike on data or downfall will misleads the training of classifier and can increase misclassification rate. Therefore, in this paper we will work to enhance the quality of data to make it outlier free which will help to improve the true fraud detection. In this section we discuss the primary aim of the proposed work involved in this paper. The next section covers the proposed solution of this problem. The next section described the experimental results and the measured performance. Finally, the conclusion will be made and future work is also discussed.

II. PROPOSED WORK

In an ML model a suitable experimental dataset is an essential component. Therefore we have found the dataset from Kaggle which is in explained form. This dataset contains credit card transactions of European cardholders of September 2013 and has 492 fraud transactions recorded among a total of 284,807 transactions. The dataset is not containing specific attribute names due to security reasons. Additionally, the dataset values are transformed using Principle Component Analysis (PCA). Only 'Time' and 'Amount' is disclosed.

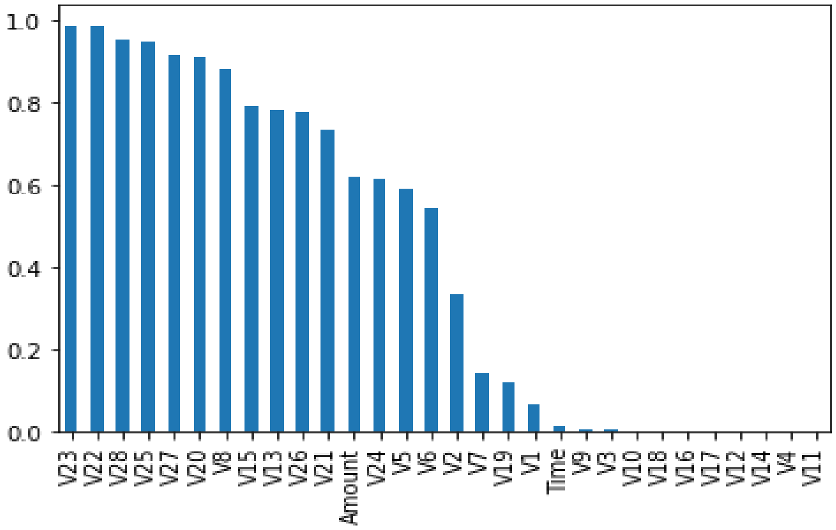


Figure 1: Shows the ranking of attributes based on p value

The transactions are labelled with a 'Class' variable and has a value of 1 for fraud and 0 for legitimate. Additionally, the dataset has a total of 30 attributes. Therefore, we performed chi-square test between the classes and the dataset attributes. The chi-square test provides two components chi-square score and p values. The sorted p values-based attribute ranking is demonstrated in figure 1.

According to the p values the attributes 'V23', 'V22', 'V28', 'V25', and 'Amount' has been selected for further experiment. Additionally, we have removed the less significant attributes, and the remaining 5 attributes and a class attribute is going to be used. Let this remaining credit card data set is D, which have a set of finite attributes such that $D = \{D_1, D_2, \dots, D_n\}$ with a decisional attribute class a C. But the dataset may contain missing value therefore a new data set D_a is generated by replacing the frequent value of the attribute. The following process is implemented as described in table 1.

Table 1: Missing Value Handling

```
Input: Dataset $D_a$
Output: clean Dataset $D_c$

Process:
1. $[col, row] = ReadDataset(D_a)$
2. $for(i = 1; i \leq col; i++)$
      a. $f = GetFrequent(D_a[i])$
      b. $for(j = 1; j \leq row; j++)$
            i. $if\ D_a[i][j] == null\ ||\ D_a[i][j] = NaN$
                  1. $D_a[i][j] = f$
            ii. End if
      c. End for
      d. $D_c.Append(D_a[i])$
3. End for
4. Return $D_c$
```

The dataset D_a is first read to get the dimension of data the total number of rows and columns are calculated first. Now for each attribute we calculate the most frequent symbol or value. Now using these frequently identified values we replace the NaN and Null value. After replacing all the missing values we found a new refined dataset D_c. But, due to PCA-based transformed values, the data has "+" positive as well as negative "-" values. Therefore, we scale the entire dataset values between 0-1. In this context, we utilized min-max normalization. The steps used for normalizing the dataset are given in table 2.

Table 2: Dataset Normalization

```
Input: dataset $D_c$
Output: normalized dataset $D_{norm}$

Process:
1. $[col\ row] = readDataset(D_c)$
2. $for(i = 1; i \leq col; i++)$
      a. $max = findMax(D_c[i])$
      b. $min = findMin(D_c[i])$
      c. $for(j = 1; j \leq row; j++)$
            i. $val = D_c[i][j]$
            ii. $newVal = \frac{val-min}{max-min}$
            iii. $D_{norm}[i][j] = newVal$
      d. End for
```

```
3. End for
4. Return $D_{norm}$
```

According to the given process the dimension of D_c additionally values are normalized for creating a new dataset D_{norm}. Now in order to increase separability of dataset we need to find the overlapped attributes. The overlapping attributes may negatively impact on classification accuracy. In this context therefore here the Mega Trend Diffusion (MTD) is used to analyse the dataset.

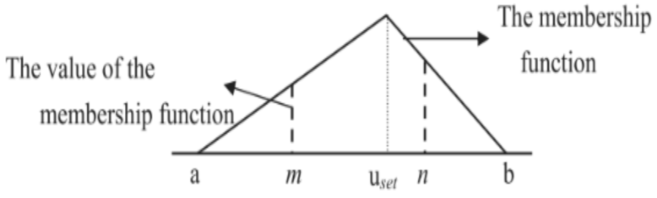


Figure 2 MTD function

The MTD is a fuzzy based technique for finding the samples which are overlapped each other. In order to explain this concept let the attribute $\mathrm{A} = \{\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_\mathrm{n}\}$, with two boundary conditions “a” and “b”. The approximation of these conditions is performed using equation (1) and (2):

$$a = u_{set} - skew_L * \sqrt{(-2) * \frac{s_a^2}{N_L * \ln(f(t))}} \dots\dots (1)$$

$$b = u_{set} - skew_U * \sqrt{(-2) * \frac{s_a^2}{N_U * \ln(f(t))}} \dots\dots (2)$$

Where,

$$u_{set} = \frac{min + max}{2} \dots\dots (3)$$

s_a^2 = Variance of attribute A_i

N_L= the number of data points smaller than u_{set}

N_U= the number of data points greater than u_{set}

$$skew_L = \frac{N_L}{N_L + N_U} \dots\dots (4)$$

And,

$$skew_U = \frac{N_U}{N_L + N_U} \dots\dots (5)$$

And, $f(t)$ is a real number greater than 0.

Next we define the MTD as membership function, which is denoted by $m(x)$ as given in equation (6).

$$m(x) = \begin{cases} \frac{x-a}{u_{set}-a}, & a \leq x \leq u_{set} \\ \frac{b-x}{u_{set}-b}, & u_{set} \leq x \leq b \\ 0, & otherwise \end{cases} \dots \dots ..(6)$$

Now, we need to calculate the overlap area to and deciding the high or low overlapped area. In our experimental dataset credit card fraud detection we have two classes F for fraud and T for legitimate. The area of MDT function of attribute A_i is β_A^i and for the same A_i for class B is given by β_B^i. Then the overlapped area of class F and T is β_O^i. Thus the rate of overlap of class F is given by β_O^i/β_A^i and for class T is β_O^i/β_B^i. Then degree of overlap is calculated by equation (7).

$$OD^i = \sqrt{\frac{\beta_O^i}{\beta_A^i} * \frac{\beta_O^i}{\beta_B^i}} \dots \dots ..(7)$$

Then a threshold for OD^i is calculated as the mean of OD=($OD^1, \dots, OD^i, \dots OD^n$). The corresponding attributes are defined as having low overlap when less than threshold T:

$$T = \frac{1}{n}\sum_{i=1}^{n} OD^1 + OD^i + OD^n \dots \dots ..(8)$$

The high and low overlap area is defined by using.

$$\begin{cases} OD^i < T, & LO \\ OD^i > T, & hO \end{cases} \dots \dots ..(9)$$

The overlapping condition shows the quality of data attributes. However, the MTD technique involves two different processes to deal with high and low overlap data. But in this experiment the less overlapped attributes are considered for performing the training. In this context only low overlapped attributes 'V23', 'V22', 'V28', 'V25', and 'Amount' has been selected for further experiment. Further for improving the data quality the outlier analysis performed. The outlier is the data points that are not in a regular manner or providing misleading patters. Basically it can be an spike or downfall in trend. Therefore, the data is used with the regression analysis for discovering the outlier points. The regression analysis is used to measure:

$$[R, R_i] = regress(D)$$

Where, R is residual of size n-by-1, R_i is n-by-2 matrix of intervals used to diagnose outliers.

If interval R_i (i, :) for observation i does not include zero, therefore residual is larger than expected in 100*(1-alpha)% is suggested as outlier. In figure 3 the outlier is demonstrated as red line plot. Additionally the green lines show the regular patterns.

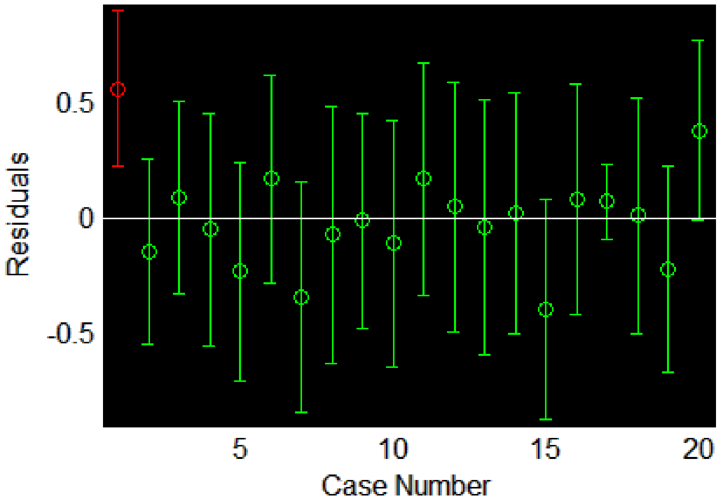


Figure 3: Outlier Detection

Therefore, outlier is a type of error located in the given data set. In order to the dataset by eliminating the outlier points we follow a step of process as given in table 3. The given algorithm accept the dimensionality reduced dataset D_{red} and generate the outlier free data P_{out}. In this context, the dataset each instance is verified using the residual values. If residual both the intervals are below or higher than zero then the data instance is considered as outlier.

Table 3 outlier detection

Input: reduced dimension of data D_{red}

Output: outlier points P_{out}

Process:

```
1. $[col\ row] = Dataread(D_{red})$
2. $[R, R_i] = regress(D_{red})$
3. $for(i = 1; i \leq row; i++)$
    a. $if(R_{i,1} \leq 0\ and\ R_{i,2} \leq 0)$
        i. $P_{out}.Add(D_{red}[i])$
    b. $else\ if\ (R_{i,1} \geq 0\ and\ R_{i,2} \geq 0)$
        i. $P_{out}.Add(D_{red}[i])$
    c. End if
4. End for
5. Return $P_{out}$
```

After preparing the final data we utilize two machine learning algorithms Xgboost and CNN to utilize for training and classification. This section provides the details about the prepared credit card fraud detection technique. The next section discusses the experimental results of the prepared system.

III. RESULTS & DISCUSSION

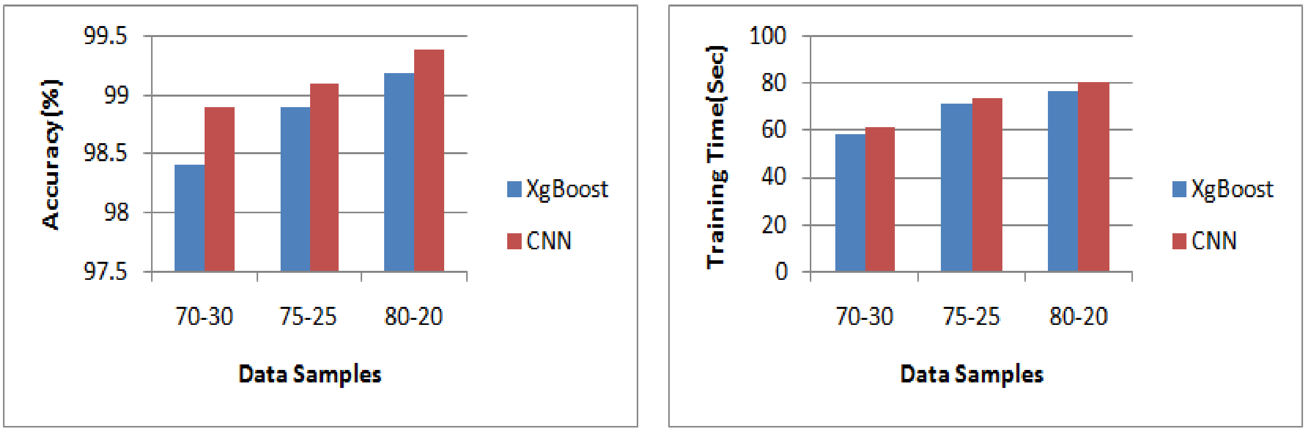
In this experiment we are investigating the data enhancement techniques therefore four different techniques are utilized. First the chi-square test is used to reduce the dimensions of the data. Next, the problem of missing values is handled. Further the MTD function is used to identify the highly overlapped attributes for reducing the dimensions more. Next, outliers are removed from the data by using regression analysis. Finally the enhanced data is utilized for training of ML algorithms and performing classification task. Finally, the performance of the machine learning algorithms is measured in terms of accuracy and training time to compare and identify the appropriate approach for credit care fraud detection.

The performance in terms of accuracy is given in figure 4. The accuracy is the ratio of correctly recognized fraud cases out of total observations are provided for recognition. It can be measured using the equation (10):

$$accuracy = \frac{correctly\ recognized}{total\ samples} \ldots \ldots . . (10)$$

Next parameter is training time, which is defined as the total amount of time taken for performing training. Thus it can be described as the time difference between training start time and training completing time. That can be calculated using equation (11):

$$Training\ time = end\ time - start\ time \ldots \ldots . . (11)$$



(A) (B)

Figure 4 shows the performance of the proposed credit card fraud detection model in terms of (A) accuracy and (B) training time

The accuracy of the proposed model using both the classification algorithms is demonstrated in figure 4(A). The X axis shows the training and testing validation samples ratio and Y axis shows the results in terms of accuracy. According to the results, the Xgboost and CNN both algorithms are providing higher classification accuracy. But the CNN provides more accurate classification.

Table 4: Performance of the proposed credit card fraud detection system

Data Samples	Accuracy (%)		Training Time (Sec)	
	XgBoost	CNN	XgBoost	CNN
70-30	98.4	98.9	58.2	61.4
75-25	98.9	99.1	71.5	73.7
80-20	99.2	99.4	77.1	80.5

Next figure 4(B) shows the training time of both the algorithm over the enhanced dataset. The training time is measured in terms of seconds. The Y axis contains the training time and the X axis shows the training and validation samples. According to the results the CNN algorithm requires more time for training as compared to XgBoost algorithm.

VI. CONCLUSIONS

The credit card frauds are one of the crucial issues for a banking company for their reputation and finance. Therefore, banking companies monitors the credit card transactions to identify the fraud transactions. In order to secure the credit card transactions ML based methods are mostly used. But, the recent techniques have suffered from noisy nature of credit card fraud dataset. Therefore, in this paper we work to reduce the noise from the dataset. The paper includes an algorithm for providing enhanced quality data in four steps.

1. Chi-square test is carried out for selecting the most informative attributes from the dataset
2. Next, an algorithm is implemented for identifying the missing values using the most frequent values in attribute
3. A method is implemented for identifying the overlapping information based on attributes. Additionally based on ranking of overlapped attributes the dimensions of data is reduced
4. Finally the residual analysis of samples was performed for identifying the outliers. The outlier detection is performed using the regression analysis.

The obtained filtered data is used with two popular machine learning techniques for training and validation. Based on the experimental results the optimization of data will help to improve the classification accuracy as well as by reducing the dataset dimensions it also reduces the time consumption.

REFERENCES

[1] Sangeeta Mittal and Shivani Tyagi, “Chapter 26: Computational Techniques for Real-Time Credit Card Fraud Detection”, Handbook of Computer Networks and Cyber Security, © Springer Nature Switzerland AG 2020

[2] G. Sasikala, M. Laavanya, B. Sathyasri, C. Supraja, V. Mahalakshmi, S. S. Sreeja Mole, Jaison Mulerikkal, S. Chidambaranathan, C. Arvind, K. Srihari, and Minilu

Dejene, “An Innovative Sensing Machine Learning Technique to Detect Credit Card Frauds in Wireless Communications”, Hindawi Wireless Communications and Mobile Computing Volume 2022, Article ID 2439205, 12 pages

[3] Ibtissam Benchaji, Samira Douzi, and Bouabid El Ouahidi, “Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks”, Journal of Advances in Information Technology Vol. 12, No. 2, May 2021

[4] Aishwarya Mujumdar, Dr. Vaidehi V, “Diabetes Prediction using Machine Learning Algorithms”, Procedia Computer Science 165 (2019) 292–299

[5] Naoufal Rtaylia, Nourddine Enneya, “Selection Features and Support Vector Machine for Credit Card Risk Identification”, Procedia Manufacturing 46 (2020) 941–948

[6] Praveen Kumar Sadineni, “Detection of Fraudulent Transactions in Credit Card using Machine Learning Algorithms”, Proceedings of the Fourth International Conference on I-SMAC, 978-1-7281-5464-0/20/$31.00 ©2020 IEEE

[7] Olawale Adepoju, Julius Wosowei, Shiwani lawte, Hemaint Jaiman, “Comparative Evaluation Of Credit Card Fraud Detection Using Machine Learning Techniques”, 2019 Global Conference for Advancement in Technology (GCAT), 978-1-7281-3694-3/$31.00 ©2019 IEEE

[8] Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, “Credit Card Fraud Detection - Machine Learning methods”, 18th International Symposium INFOTEH-JAHORINA, 20-22 March 2019, 978-1-5386-7073-6/19/$31.00 ©2019 IEEE

[9] Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar Mohammad Masum, “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset”, 8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019 Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

[10] Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya, Maheshwar Sharma, “Credit card fraud detection using Naïve Bayes model based and KNN classifier”, International Journal of Advance Research, Ideas and Innovations in Technology, Volume 4, Issue 3, 2018

[11] Imane Sadgali, Nawal Sael, Nawal Sael, “Fraud detection in credit card transaction using neural networks”, SCA2019, October 2–4, 2019, CASABLANCA, Morocco, © 2019 Association for Computing Machinery

[12] Tzu-Hsuan Lin and Jehn-Ruey Jiang, “Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest”, Mathematics 2021, 9, 2683.

[13] Dileep M R, Navaneeth A V, Abhishek M, “A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms”, Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). 978-1-6654-1960-4/21/$31.00 ©2021 IEEE

[14] P. Shanmugapriya, R. Shupraja, V. Madhumitha, “Credit Card Fraud Detection System Using CNN”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 10 Issue III Mar 2022