International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I6P104

Volume 68 Issue 6, 22-26, May 2020
© 2020 Seventh Sense Research Group®

Review Article

Review on Credit Card Fraud Detection using Machine Learning Algorithms

Pooja[1], Ashlesha[2]

[1] Research Scholar (Network engineering), J.C. Bose University of Science and Technology
YMCA at Faridabad, Haryana - India

[2] Assistant Professor(Computer engineering department), J.C Bose University of Science and Technology, YMCA at Faridabad – Haryana – India.

Received Date: 28 April 2020
Revised Date: 08 June 2020
Accepted Date: 09 June 2020

***Abstract** - The advancement of new technologies and the fast-growing of technological development have generated new possibilities as well as imposed new challenges. Fraud, the biggest challenge for business and organization, emerge with new technologies to take new and distinctive forms that are hidden and tougher to identify than the conventional forms of this crime. Credit card frauds also grow up along with growth in technology. It also noticed that financial fraud is extremely growing in the global communication improvement. It is being admitted every year that the loss because of these types of fraudulent activities is billions of dollars. These activities are performed so gracefully that they look similar to original transactions. Simply using of pattern matching technique and simple method is really not useful for detecting these fraudulent activities. A well planned and systematic method has become a need for all businesses and organizations to minimize chaos and carry it out in place. Several techniques have been evolved based on Artificial intelligence, Machine learning, Data mining, Genetic programming, Fuzzy logic etc.. for detecting credit card fraudulent activities. Besides this technique, the K-Nearest Neighbour algorithm and outlier detection methods are implemented to optimize the best solution for the fraud detection problem. These techniques proved to minimize the false alarm rates and increase the fraud detection rate.*

***Keywords** – Credit Card, Fraud Detection,* Machine Learning, Logistic Regression, K-Nearest Neighbour.

I. INTRODUCTION

Credit card fraud detection is a way of recording the selling and buying behaviour of the customer during a certain period. A credit card is just a plastic card given to customers As a payment system. Cardholders can buy Goods and Services on the basis of a promise to pay for using these Goods and Services. Security of credit cards depends upon the physical security and on the privacy of credit card Numbers. Globalization And due to increasing use of the internet for Online Shopping has to Result in a substantial increase in credit card transactions around the world. Due to the growth in the use of credit card transactions, there is also very growth in fraud activities. In a given transaction, credit card fraud is a term used for theft and fraud observed by using credit cards as a source of fraud funds. Regular use of credit card transactions for obtaining Goods and Services assisting online or card swiping method leading to continuous growth in the online transaction by use of credit and debit card increasing to the world of relaxing expenses. This fraud of credit cards caused greater damage to the customers and service providers. This is becoming very bad in the coming future. This person finds and adapt to changes in technology and find new and easy clever ways to do these fraudulent activities. Fraud due to these activities are very dangerous and risky. The very smart fraudster creates many identities and does online transactions without being caught. It is very hard to detect these fraud activities as these activities look real, and datasets are not available easily. The bank's owner and service provider are not interested in sharing the dataset for experiments.

Fraud can be interpreted as fraud committed to presenting the financial statements of a company. Fraud (fraud) is intentional fraud that gives a profit to the company and causes losses to the company. Fraud committed by several parties is one of the most interesting. In general, there are three things that encourage the transfer of fraud (pressure), opportunities (opportunities) and justification for the actions taken (rationalization). Encouragement (pressure) is an impetus arising from a desire to get a better life. Luxury supported by poor economic conditions also encourages the environment for a luxurious lifestyle. Opportunity (opportunity) is a deception factor used by his weak party Rationalization (rationalization) is a reason made by the party who committed fraud.

The types of fraud can be grouped into 3, namely:

- Employee fraud (employee fraud), namely fraud committed by employees in a work organization.

- Management fraud is fraud committed by management using financial statements or financial transactions as fraud.
- Computer Fraud (Computer Fraud) is fraud committed to recording computers that contain operational or bookkeeping records in a company.

A. *Types of credit card fraud*

a) One is application fraud, where an individual will forge an application in order to receive a credit card. They give incorrect information about /her financial status and receive a credit card.

b) Second is assumed identity, where an individual assumes someone identity and forge a name with a temporary address.

c) The third is financial fraud, where individuals apply for a credit card with her/ own name, but information related to them is false. This happens when an individual wants more than the credit that they currently have.

d) The fourth type is skimming technology, where only the purpose is to collect and store information on a credit card. Magnetic card skimming is a small handheld device used for this purpose.

e) The last type is never getting an issue where individuals steal the card while the card is in use. This type uses holder mail for stealing the card.

B. *Data mining Fraud Detection Technique*

a) Supervised Learning for Fraud Detection

In this method, all recorded datasets are classified into fraudulent and non-fraudulent earlier. Machine classify records in accordance with training given. This method only identified the fraud that has already happened, and the system trained already about these. Few of supervised techniques used are

1. Logistic regression is one of the most popular classification algorithms in machine learning. The logistic regression model describes the relationship between predictors that can be continuous, binary, and categorical. The dependent variable can be binary. Based on some predictors, we predict whether something will happen or not. We estimate the probability of belonging to each category for a given set of predictors.

2. Naive Bayes is one of the supervised learning algorithms in which there are no dependencies between attributes. It's based on the Bayes theorem. Depending on the type of distribution, there are the following algorithms: Gaussian distribution, Multinomial distribution, Bernoulli distribution. In this research, Bernoulli distribution is used for detecting fraud transactions.

3. Random forest is an algorithm that can be used in both classification and regression problems. It consists of many decision trees. This algorithm gives better results when there is a higher number of trees in the forest and prevents the model from overfitting. Each decision tree in the forest gives some results. These results are merged together in order to get a more accurate and stable prediction

4. Multilayer perceptron is feeding forward artificial neural network that consists of minimum 3 layers of nodes: input layer, hidden layer and output layer. Each node uses an activation function. The activation function calculates the weighted sum of its inputs and adds bias. This allows us to decide which neuron should be removed and not considered in outside connections.

5. KNN is the K-nearest neighbour algorithm that is used largely in fraud detection systems using supervised learning techniques. In KNN, newly arrived data classified depend upon the KNN category. This method was first utilized by Aha, Albert and Kibler in the year 1991. The outcome of KNN is based upon three factors:

a). The distance metric is used to decide the nearest neighbours.

b). The distance rule that is used for the classification from K- nearest neighbour.

c). The number of neighbours considered to classify the new sample.

6. C4.5 algorithm is one of the classification algorithms introduced by J. Ross Quinlan (1996) as an improved version of ID3. C4.5 algorithm is the development of the Decision Tree classification tree or decision tree. Primarily, the selection of the breaking point (node) in this algorithm is based on the Gain calculation to induce the tree to be formed. In the classification process of the C4.5 algorithm, there are destination variables that are usually grouped with certainty. Next, a decision tree model will be formed by calculating the probability of each record for each feature. Every data to be tested using the C4.5 algorithm needs to be divided into training data and testing data for each feature and label.

b) *Unsupervised Learning for FraudDetection*

This technique only identify the likelihood of a few records being more fraudulent than other records without any assurance. There is no classification earlier. All records are separately handled.

The KMeans algorithm is the best-known partitioned classification algorithm which is a simple method for estimating the average (vector) of a K groupset. Kmeans is the most widely used among all clustering algorithms because of its efficiency and simplicity. KMeans is a well-known and widely used grouping algorithm. KMeans is one of the simplest grouping algorithms in machine learning that can be used to automatically recognize groups of similar objects in training data.

II. LITERATURE REVIEW

N. Malini, Dr M. Pushpa, the author of a research paper[1], says that K-Nearest Neighbors (KNN) and outlier detection techniques are very efficient in fraud detection. This technique has proven useful in minimizing false alarm rates and increasing fraud detection rates. The author takes a new object set and firstly take one nearest neighbour and then increase the number of nearest neighbour one each time. The author takes a simple example of a positive and negative point. By taking one neighbour vote gone to positive sign and by taking two neighbours, both positive and negative points have equal voting. By taking 5 neighbours, the author finds that voting for a positive sign is more than a negative sign. So voting goes to the positive class, and the new object goes to the positive class. The author suggests taking a value of k always an odd number.

J.O. Awoyemi et al., the authors of the research paper [2], finds KNN algorithm performed well where the authors tested and compared it with other classical algorithms used for credit card fraud detection. The author does the comparative performance of Naïv Bayes, K Nearest neighbour and Logistic regression models in the binary classification of imbalanced credit card fraud data. Three classifiers based on different machine learning techniques (Naïve Bayes, K-nearest neighbours and Logistic Regression) are trained on the real-life of credit card transactions data and their performances on credit card fraud detection evaluated and compared based on several relevant metrics. The performances of the three classifiers are examined on the two sets of data distributions using accuracy, sensitivity, specificity, precision, balanced classification rate and Matthews Correlation coefficient metrics. The dataset contains 284,807 transactions, where 492 transactions were frauds, and the rest were genuine. Considering the numbers, we can see that this dataset is highly imbalanced, where only 0.173% of transactions are labelled as frauds.

Z. Kazemi, H. Zarrabithe author of the research paper[3], made a comparison of various deep learning techniques. The author compares certain machine learning algorithms for the detection of fraudulent transactions. Hence comparison was made, and it was found that the Random Forest algorithm gives the best results, i.e. best classifies whether transactions are fraud or not. This was established using different metrics, such as recall, accuracy, and precision.

N. Kalaiselvi, S. Rajalakshmi, J. Padmavathi Authors of paper[4] used neural networks in order to demonstrate improvement in results when ensemble techniques are used. For analysis, the author used a sample set of 5,850 fraud transactions and 542,858 legal transactions, ordered by their timestamps. It should be noted that the mining algorithm has a high runtime complexity. So author used only 30,000 of the legal transactions. The resulting values for the confidence were compared to the whole set of transactions. The author uses only fraud occurrences of .1%; the simple constant diagnosis "transactions is no fraud" will have a success rate of99.9%. To compete with this trivial diagnosis, the task of diagnosing a transaction is not easy-to-do. If the author use only the analogue data, all transactions patterns characterized by n symbolic and m analogue features are projected from the n+m dimensional space into the m dimensional space. Generally, this results in overlapping classes and, therefore, in diagnostic success far worse than 99.9%.

Mrs C. Navamani, M. Phil, S.Krishnan author of the research paper [5], find outlier detection techniques are very efficient in fraud detection. This technique has proven useful in minimizing false alarm rates and increasing fraud detection rates. The author studies different outlier detection techniques. Apply this technique for fraud detection. Credit card fraud can be solved using these outlier methods. These methods are useful in detecting fraudulent activities.

F. Ghobadi, M.Rohani the author of the research paper[6], make a comparison of neural network and various ensembles techniques and finds neural networks are better performance in comparison of ensembles techniques. Three classifier models based on Neural network, k-nearest neighbour and logistic regression are developed. To evaluate these models, 70% of the dataset is used for training, while 30% is set aside for validating and testing. Accuracy, sensitivity, specificity, precision, Matthews correlation coefficient (MCC) and balanced classification rate are used to evaluate the performance of the three classifiers. The accuracy of the classifiers for the original 0.172:99.828 dataset distribution, the sampled 10:90 and 34:66 distributions are presented in Tables 1, 2 and 3, respectively. An observation of the metric tables shows that there is a significant improvement from the sampled dataset distribution of 10:90 to 34:66 for accuracy, sensitivity, specificity, Matthews correlation coefficient and balance classification rate of the classifiers. This shows that a hybrid sampling(under-sampling and over-sampling) on a highly imbalanced dataset greatly improves the performance of binary classification. The true positive, true negative, false positive and false negative rates of the classifiers in each set of unsampled and sampled data distribution is shown in Tables 4 5and6. Logistic regression is the only technique that did not show better improvement in false-negative rates from the 10:90 to 34:66 data distribution. However, it showed the overall best performance in the un-sampled distribution.

Venkata Ratnam Ganji et al. [7]the author use concept of data stream outlier detection algorithm, which is based on anti k nearest neighbour for credit cards fraud detection, whereas traditional methods need to scan the dataset many times to find fraudulent transactions, which is not suitable for data stream surroundings. This method makes it easier to stop fraudulent transaction that happens with lost and stolen credit card .validation check and detects errors in a sequence of numbers, which also helps to detect valid and invalid numbers easily.

Abhinav Srivastava, the author of the research paper[8], uses the ranges of the transaction amount as an attribute in the HMM. The author has suggested a method for finding the spending profile of cardholders. It is also discussed how the HMM can identify fraudulent transactions. The simulation results show the advantages of using HMM, and learning the profile of the cardholder plays an important role in analyzing fraudulent cases. The result also shows that 80% of the results are accurate, and the system is scalable for large data set as well.

Divya. Iyer et al. [9] the author uses Hidden Markov Model (HMM) to detect credit card transaction frauds. The training set is tuned with the normal behaviour of the cardholder. So if a credit card transaction is rejected by the trained HMM, then that transaction is said to be fraudulent. Care is to be taken that valid and genuine transactions are not considered fraud. The author also compares various methods with the proposed methods to prove that HMM is much preferred than the other methods.

K.RamaKalyani et al. [10] create test data through which fraudulent activities are detected. This algorithm is also called an optimization technique based on genetic and natural selection in great computational problems. The author proposes a method to detect credit card fraud, and the results are validated using the principles of this algorithm. The purpose of detecting fraud cases is to declare it to the client and the service provider.

Renu et al. [11] proposed a fraud detection method that involves monitoring the activities of populations to observe and predict undesirable behaviour. Undesirable behaviour is a set of several habits like intrusion, fraud, delinquency and default.

K.Swapna, Prof. M.S. Prasad Babu [12] depicted that, Design a liver diagnosis system automatically to detect early and accurately to reduce deaths caused by liver disease and analyze data sets to understand the system to design a liver diagnosis system automatically to detect early and accurately to reduce deaths caused by liver disease and analyze data set to understand the use system.

Heta Naik [13] depicted that, The increase in online transactions is directly proportional to the increasing number of frauds. In this paper, various algorithms such as K-Nearest Neighbor, Random Tree, AdaBoost and Logistic Regression are some challenges which include distinguishing between normal transactions and fraud that seem very similar to each other. The parameters to detect these transactions are Time, number and Frequency of Transactions. In this paper, four different KNN algorithms, AdaBoost, Random tree and Logistic Regression, are compared for fraud detection mechanisms. Logistic regression is better compared to other algorithms. This model is used for unbalanced credit card fraud data. All of these algorithms do not apply to fraud detection at the time of the transaction.

Yezheng Liu et al. [14] depicted that, Approach outlier detection as a binary classification problem by taking potential outliers from a uniform reference distribution. However, due to the scarcity of data in high-dimensional space, a number of potential outliers may fail to provide information to help the classifier draw a line that can effectively separate outliers from normal data. To overcome this, propose a Single Objective Objective Active Target (SO-GAAL) Active Learning method for outlier detection, which can directly generate potential informative outliers. Proposed a new SOGAAL outlier detection algorithm, which can directly produce potentially informative outliers, to overcome the lack of information caused by curse dimensions. Extending the GAAL structure of a single generator (SO-GAAL) to several generators with different purposes (MO-GAAL) to prevent generators from falling into a collapsing problem mode. Compared to some sophisticated outlier detection methods, MO-GAAL achieves the best average rating in real-world datasets and shows strong robustness for various parameters. In addition, MO-GAAL can easily handle various types of clusters and high irrelevant variables.

K.T.Divya, N.Senthil Kumaran [15] depicted that, The outlier detection approach is based on distance learning for category attributes (DILCA), a distance learning framework is introduced. The key intuition of DILCA is that the distance between two categorical attribute values can be determined in a way where they occur together with other attribute values in the data set. The classic KNN produces superior data utility but raises a higher computational overhead. In addition, dimension reduction techniques used in the occupational health dataset are used.

Table 1.

Fraud Detection Technique	**Advantage**	**Disadvantage**
Logistic Regression	Work well with linear data and detect fraud by creating a probabilistic formula for classification	Can not handle non-linear data can not detect fraud at the time of transaction
Decision Tree	Can apply on linear as well as on non-linear data	Algo very complex even a small change can change the whole tree structure
Hidden Markov Model	It can detect fraudulent activities at the time of transactions and reduce the false-positive ratio	It can not detect fraud at some initial transactions
Artificial neural network	Can detect fraud activity at the time of online transaction	The number of the parameter is to be set before training started no certain method till now for deciding optimal topology for a particular problem network working depends upon the interconnection of neurons
K-nearest neighbour	No requirement of predictive model before classification	Can not detect fraud at the time of the transaction, and accuracy depends on the distance measured

III. CONCLUSION

Credit card fraud has become very large these days. To progress safety measures of the monetary transaction systems in a habitual and effectual way, structuring a precise and well-organized credit card scam detection system is one of the essential functions for money transactions. By performing oversampling and extracting the principal direction of the data, we can use our KNN method to determine the anomaly of the target instance. Hence the KNN method can suit for detecting fraud with the limitation of memory. In the meantime, the outlier detection mechanism helps to detect credit card fraud using less memory and computation requirements. Especially outlier detection works fast and well on large online datasets. But compared with power methods and other known anomaly detection methods, experimental results prove that the KNN method is accurate and efficient

REFERENCES

[1] N. Malini, Dr M. Pushpa, Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection Advances in Electrical, Electronics, Information, Communication and BioInformatics(AEEICB), Third International Conference on pp. 255258. IEEE. (2017).

[2] J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis, Computing Networking and Informatics (ICCNI), International Conference on IEEE.(2017) 1-9.

[3] Z. Kazemi, H. Zarrabi, Using deep networks for fraud detection in the credit card transactions, Knowledge-Based Engineering and Innovation (KBEI), IEEE 4th International Conference on (2017) 630-633.

[4] N. Kalaiselvi, S. Rajalakshmi, J. Padmavathi, Credit card Fraud Detection using Learning to Rank Approach, 2018 Internat International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC) ional conference on computation of power, energy, Information and Communication (ICCPEIC) (2018)191-196.

[5] Mrs C. Navamani, M. Phil, S. Krishnan, Credit Card Nearest Neighbor Based Outlier Detection Techniques, International Journal of Computer Techniques 5 (2) (2018) .

[6] F. Ghobadi, M. Rohani, Cost Sensitive Modeling of Credit Card Fraud using Neural Network strategy, Signal Processing and Intelligent Systems (ICSPIS), International Conference of IEEE(2016) 1-5

[7] Venkaata Ratnam Ganji, credit card Fraud Detection using Anti-K Nearest Neighbour Algorithm, Internationnal Journal on Computer Science and Engineering(IJCSE) 4(6) (2012) 1035-1039

[8] Abhinav Srivastava, Amlan Kundu, Shamik Sural, and Arun K. Majumdar Credit Card Fraud Detection Using Hidden Markov Model 5(1) (2008).

[9] Divya.Iyer, Arti Mohanpurkar, Sneha Janardhan, Dhanashree Rathod, Amruta Sardeshmukh Credit card Fraud Detection using Hidden Markov model 978-14673- 01268/11/$26.00_c IEEE (2011).

[10] K.RamaKalyani, D.UmaDevi Fraud Detection of Credit Card Payment System by Genetic Algorithm 3(7) (2012)

[11] Renu, Suman. Analysis on Credit Card Fraud Detection Methods. International Journal of Computer Trends and Technology (IJCTT) 8(1) (2014) 45-51. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.

[12] Swapna K, and Babu M S P International Journal of Electrical & Computer Sciences IJECS-IJENS A Framework for Outlier Detection Using Improved Bisecting KMeans Clustering Algorithm. 17 0812 (2017)

[13] Naik H International Journal for Research in Applied Science & Engineering Technology (IJRASET) Credit Card Fraud Detection for Online Banking Transactions 6 453457 (2018)

[14] Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang and Xiangnan He / Generative Adversarial Active Learning for Unsupervised Outlier Detection (2019)

[15] Divya K T, Kumaran N S International Research Journal of Engineering and Technology (IRJET) Improved Outlier Detection Using Classic Knn Algorithm 3 892898(2016).

[16] Pooja Bhati, Manoj Sharma, Credit Card Number Fraud Detection Using K-Means with Hidden Markov Method, SSRG International Journal of Mobile Computing and Application, 2(2)(2015).