IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10320

Credit Card Fraud Detection Using Machine Learning

Mr.G.Deeban Chakkarawarthi[1], Irfana K V[2], Naseela K T[3], Akshay Sivaraj[4]

Department of Computer Science and Engineering, JCT College of Engineering and Technology
Coimbatore, TamilNadu, India[1-4]

Abstract: Credit card has an important role in our day- to-day lives. It is now being widely used all over the world for transactions, irrespective of the geographical boundaries People use credit card for their purchase, it allows them to pay it later. Credit card fraud happens when someone steals the information or loses the card. Criminals may be using technologies such as Trojan or Phishing to get card details. Therefore, an effective fraud detection method is important since it can identify a fraud in time when a criminal uses a stolen card to consume. One method is to make full use of the historical transaction data including normal transactions and fraud ones to obtain normal/fraud behavior features based on machine learning techniques, and then utilize these features to check if a transaction is fraud or not. In this paper, Machine Learning algorithm is used to train the behavior features of normal and abnormal transactions. We implement this using Random forest machine learning algorithm in OpenCV and analyze the performance on credit fraud detection.

Keywords: Credit card fraud, Machine learning, Random Forest, openCV

I.INTRODUCTION

The advent of technology, in the form of the credit card, brought convenience and made life simpler. While credit cards have made life easy, they have also managed to make life easy for the crooks. While it made fraudsters to guzzle away money that is not truly theirs. There is a growing trend of transaction frauds resulting in a great losses of money every year. It is estimated that losses are increased yearly at double digit rates by 2020. Since the physical card is no needed in the online transaction environment and the card's information is enough to complete a payment, it is easier to conduct a fraud than before. Transaction fraud has become a top barrier to the development of e-commerce and has a dramatic influence on the economy. Hence, fraud detection is essential and necessary.

The key objective of credit card fraud detection system is to identify suspicious events and report them to an analyst while letting normal transactions be automatically processed. For years, financial institutions have been entrusting this task to rule-based systems that employ rule sets written by experts. But now they increasingly turn to a machine learning approach, as it can bring significant improvements to the process.
There are two kinds of methods for fraud detection, misuse detection and anomaly detection. Misuse detection uses classification methods to determine whether an incoming transaction is fraud or not. Usually, such an approach has to know about the existing types of fraud to make models by learning the various fraud patterns. Anomaly detection is to build the profile of normal transaction behavior of a cardholder based on his/her historical transaction data, and decide a newly transaction as a potential fraud if it deviates from the normal transaction behavior. However, an anomaly detection method needs enough successive sample data to characterize the normal transaction behavior of a cardholder.

This paper is about misuse method. We use Random forest to train the normal and fraud behavior features. Random Forest is a classification algorithm that is comprised of many Decision Trees. Each tree has nodes with conditions, which define the final decision based on the highest value. The Random Forest algorithm for fraud detection and prevention has two cardinal factors that make it good at predicting things. The first one is randomness, meaning that the rows and columns of data are chosen randomly from the dataset and fit into different Decision Trees. The second factor is diversity, meaning that there's a forest of trees that contribute to the final decision instead of just one decision tree. The biggest advantage here is that this diversity decreases the chance of model overfitting, while the bias remains the same.

Some of the currently used approaches are:

- Markov model
- Artificial Neural Network
- Cost sensitive decision tree
- Support vector machine
- Neuro-fuzzy

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10320

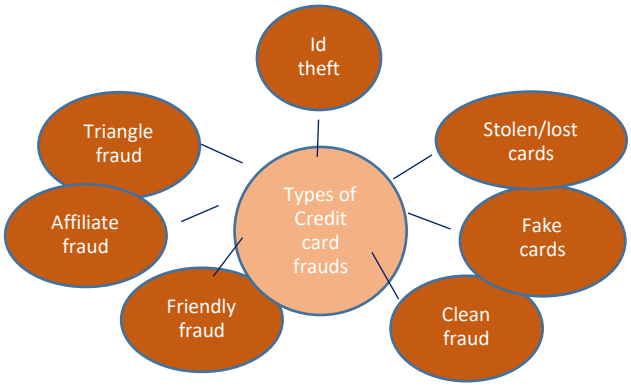


Fig.1: Types of credit card frauds

II.RELATED WORK

A comprehensive understanding of fraud detection technologies can be helpful for us to solve the problem of credit card fraud. The work in [1] provides the most common methods of fraud along with their detection methods and reviewed recent findings in this field. It also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. They focused on analysing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm on the PCA transformed Credit Card Transaction data.

A similar research domain was presented by Sikhakolli VenkatavisalakshiseshsaiYasaswi, Sahayasakila.V, D. Kavya Monisha and Aishwarya where they used Smote technique and whale optimization algorithm[2]. The Smote technique is used to solve Class imbalance problem. The Whale optimization algorithm comprises mainly of three operators which are used to stimulate the search for prey, encircling prey and bubble-net scratch around the behaviour of humpback whales.

Fraud detection which is based on neural networks is also one of the most popular methods. S. Ghosh et al. [9] have implemented a neural network-based system to detect fraudulent credit card transactions. This model is able to learn from the past because it is formed by all types of transactions over a period of time. Azeem Ush Shan Khan et al. [7] have developed a system on a neural network when trained with simulated annealing algorithm. But the problem is that the user's activity is different in each transaction which makes it difficult to form any ANN. Next is the clustering technique [10]which analyzes the spending behavior of the credit card user to prevent and detect fraud. In other words, when a transaction violates a certain account behavior in an unusual manner, an alarm is triggered and the transaction is declared fraudulent. These behaviors can be seen through the unusual amount of money that the cardholder isn't used on using at once, the expenses and items purchased and many others.

Another model has been proposed by K.R. Seeja and Masoumeh Zareapoor [8]falling into the same topic of credit card fraud detection. This model detects fraud from an unidentified and imbalanced credit card transactions dataset. In order to discover if the incoming transactions of the clients are legal or illegal, a parallel algorithm is proposed. Another architecture which deals with the same issue uses HDFS to store and rapidly access the logs is Anushree and Kalyani Phumamdikar's[5]. They suggested in their work that using a Bayes minimum risk classifier sheds light on better fraud detection results in the sense of higher savings.

III. METHODOLOGY

The approach that this paper proposes, uses the machine learning algorithms called Random Forest. The main objective of this algorithm is to classify the collected data [here transactional data] as normal and abnormal. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. Machine learning algorithms are employed to analyze all the authorized transactions and report the suspicious ones.

1.Data Collection and Analysis

In this project, we have to collect the data from the bank. But Bank does not provide transactional datas. So we collected datas from internet which already uploaded by some companies. The data maybe include NaN values or missing fields. So that we analyze these datas. data analysis is a process in which data from one or more sources is cleaned, transformed and enriched to improve the quality of data prior to its use. The collected data were then pre-processed to fill the missing data and made compatible for further processing.

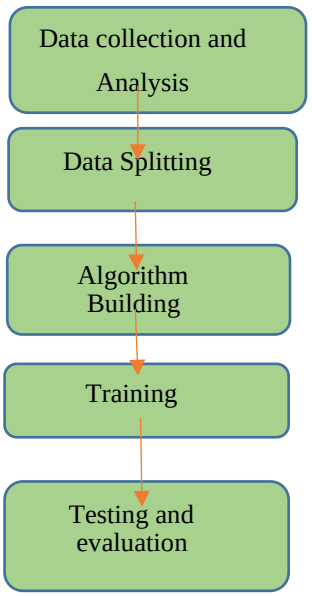
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10320



2.Data Splitting

Data splitting is the act of partitioning available data into. two portions, usually for cross-validatory purposes. One. portion of the data is used to develop a predictive model. and the other to evaluate the model's performance.ie Training and Testing. Our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to training set and therefore the remaining 20% to test set.

3.Algorithm Building

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. The capacity of random forest not only depends on the strength of individual tree but also the correlation between different trees. The stronger the strength of single tree and the less the correlation of different tress, the better the performance of random forest. The variation of trees comes from their randomness which involves bootstrapped samples and randomly selects a subset of data attributes. Although there possibly exist some mislabeled instances in our dataset, random forest is still robust to noise and outliers.

Random-tree-based Random forest

A base classifier of random forest I, which is a simple implement of decision tree, is called a random tree [22]. The training set of each tree is a collection of bootstrapped samples selected randomly from the standard training set with replacement. At each internal node, it randomly selects a subset of attributes and computes the centers of different classes of the data in current node. The centers of class 0 and 1 are denoted as leftCenter and rightCenter, respectively. The kth element of a center is computed based on the following equations.

$$\text{lef tCenter}[k] = 1/n \sum_{i=1}^{n} x_{ik} I(y = 0) \quad (1)$$

$$\text{rightCenter}[k] = 1/n \sum_{i=1}^{n} x_{ik} I(y = 1) \quad (2)$$

where I (y = 0) and I (y = 1) are the dictator functions. At the current node, each record of the dataset is allocated to the corresponding class according to the Manhattan distance between the record and the center as shown in (3).

$$\text{Distance(center,record)} = \sum_{i \in sub} \text{center}[i] - \text{record}[i] \quad (3)$$

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10320

Note, sub is the subset of attributes randomly selected from X whose size is the square root of m = |X|. Each tree grows fully without pruning.

Algorithm:

Input: Dataset D and the number of trees NT

Output: A random forest

For i = 1 to NT:

1) Draw a bootstrap sample Di from the training set D whose size is n.

2) Construct a binary tree of the bootstrapped data recursively from root node. Repeatedly perform the following steps until all records of current node belong to a class.

a) Randomly select a subset of √m attributes.

b) For j = 1 to √m:

i) Compute lef tCenter[j] and rightCenter[j].

c) For k = 1 to |Dic|:

i) Compute the Manhattan distance dL_k and dR_k between the $record_k$ and each center.

ii) if $dL_k \leq dR_k$

Allocate recordk to the left child of the current node.

else

Allocate recordk to the right child of the current node.

d) Split the node into a left child and a right child.

where D_{ic} is the subset of D_i in the current node.

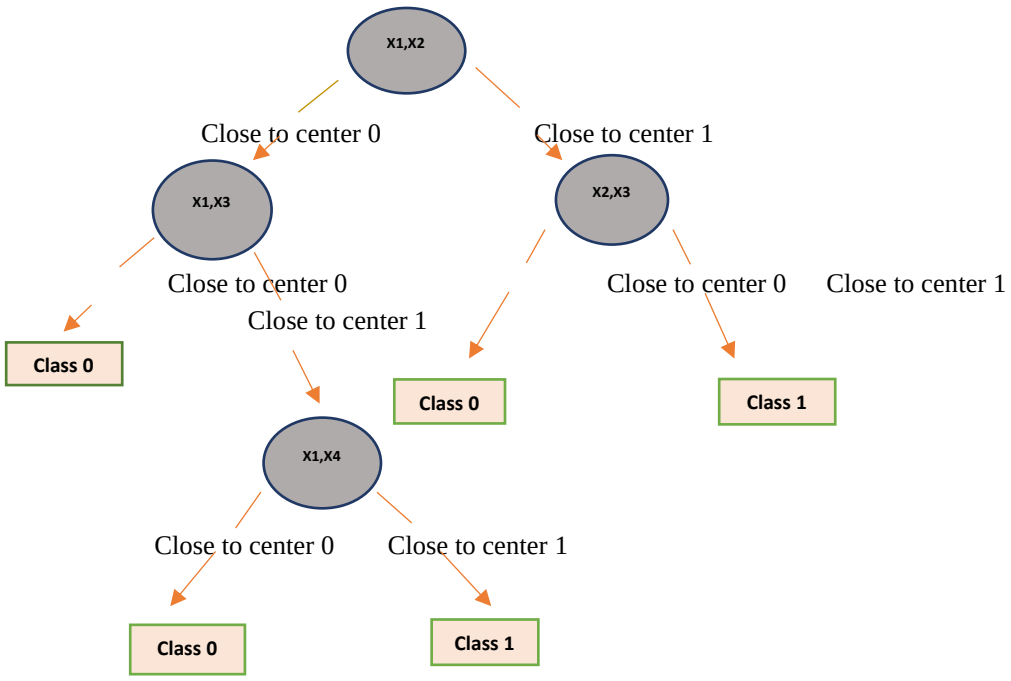


Fig. 3: Illustration of Random Tree

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10320

A simple example of random tree is shown in figure. The internal nodes are represented by circles. The variables in a circle are attributes randomly chosen from X = {x1, x2, x3, x4}. The decisions are made according to their values. Each terminal node is represented by a rectangle and corresponds to a class. The number in a terminal node represents which class the node belongs to.

Advantages:

- It reduces overfitting in decision trees and helps to improve the accuracy
- It is flexible to both classification and regression problems
- It automates missing values present in the data
- It works well with both categorical and continuous values
- Normalising of data is not required as it uses a rule-based approach.

4.Training

Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples.

5.Testing and Evaluation

Testing these datas to find the amount of normal and abnormal transaction. accuracy rate is not enough to measure the performance of a random forest model when the data is significantly imbalanced. For instance, a default prediction of all instances into the majority class will also have a high value of accuracy. Therefore, we need to consider other measures. Positive corresponds to fraud instances and Negative corresponds to normal instances. Evaluate these datas and check its accuracy.

IV. CONCLUSION

Credit card fraud has increased exponentially in recent years. Accordingly, one of the main tasks of the financial industries is to develop an accurate and easy credit card fraud detection system. There are different categories of transaction scams as well as several methods to detect them. However, all these methods remain insufficient. This paper has examined the performance of Random forest model, which identifying the fraud transactions occurring during the transactions made by the card holder. It is an effective technique that helps implement a credit card fraud detection system.

V. FUTURE WORK

While we couldn't reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. We only using the data from the internet, its impossible to get the information from bank.

One we can change is that use large amount of data and apply multiple algorithms. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives.

REFERENCES

[1] "Credit Card Fraud Detection using Machine Learning and Data Science" published by International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 8 Issue 09, September-2019

[2] "Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm" published by International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019

[3] "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10320

[4] Credit Card Fraud Detection through Par enclitic Network Analysis By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370

[5] A. Naik, K. Phulmamdikar, S. Pradhan, and S. Thorat, "Real Time Credit Card Transaction Analysis," vol. 1, no. 11, pp. 1663–1666, 2016

[6] Y. Gmbh and K. G. Co, "Global online payment methods:Full year 2016," Tech. Rep., 3 2016

[7] A. Khan, N. Akhtar, and M. Qureshi, "Real-Time Credit-Card Fraud Detection using Artificial Neural Network Tuned by Simulated Annealing Algorithm," nt. Conf. Recent Trends, 2014

[8] K. R. Seeja and M. Zareapoor, "FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining," Sci. World J., vol. 2014, pp. 1–10, 2014

[9] K. K. Tripathi and M. A. Pavaskar, "Survey on Credit Card Fraud Detection Methods," Int. J. Emerg. Technol. Adv. Eng., vol. 2, no. 11, p. 721, 2012

[10] R. J. Bolton, D. J. Hand, and D. J. H, "Unsupervised Profiling Methods for Fraud Detection," Proc. Credit Scoring Credit Control VII, pp. 5–7, 2001