IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 2, Ser. IV (Mar - Apr 2019), PP 72-76
www.iosrjournals.org

Credit Card Fraud Detection Using Hidden Markov Model and Naïve Bayes

Joyce Lemos[1], ShubhamPatil [2], SahilSave[3], Hardik Pise[4]
[1](St. John College of Engineering & Management, India)
[2](St. John College of Engineering & Management, India)
[3](St. John College of Engineering & Management, India)
[4](St. John College of Engineering & Management, India)

Abstract: *In the past few years uses of electronic commerce technology and credit card has increased dramatically. As a credit card becomes the most popular way of paying there are cases of fraud associated with regular purchases. An HMM is initially trained with the usual behavior of a cardholder. If someone's Credit card transactions are not accepted by trained HMM With sufficiently high probability, it is considered Fraudulent. At the same time, we try to ensure that the real Transaction is not denied using (hybrid model).In this paper we presents the HMM improved Naïve Bayes method for Fraud Detection of Credit Card. Experimental results illustrate that both classifiers work differently for the same dataset. The purpose is to enhance the accuracy and enhance the flexibility of the algorithm.*

Keywords: *Credit card, Fraud, Hidden Markov Model(HMM), Naïve Bayes.*

Date of Submission: 16-04-2019 Date of acceptance: 01-05-2019

I. Introduction

In day to day life credit cards are used for purchasing goods and services by the help of the virtual card for online transaction or physical card for offline transaction in physical transaction credit card will insert into payment machine at merchant shop to purchase goods.Online transactions are increasing day by day. It is very helpful in daily routine life. When online transactions are increasing, fraud transactions also increases [1]. The only way to detect this kind of fraud is to analyse the spending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on analysis of existing purchase data of card holder is promising way to reduce the rate of successful credit card frauds [2].

For every credit card, the spending profile is different, so it can figure out an inconsistency of every user profile and try to find fraudulent transaction. It keeps record of spending profile of the card holder thus analysis of purchase product of cardholder will be useful tools in fraud detection system and it is assuring the way to check fraudulent transaction, although fraud detection system does not keep record of number of purchased goods and categories. The particulars of purchased items in single transactions are generally unknown to any Credit card Fraud Detection System running either at the bank that issues credit cards to the cardholders or at the merchant site where goods is going to be purchased [3].Several techniques for detection of credit card fraud have been proposed in the last few years.

Credit-card-based purchases can be classified into two Type: 1) Physical card and 2) Virtual card. In a physical-card, cardholder physically presents his card to make a payment. In this kind of purchase, the attacker steal Credit Card if the cardholder does not realize its lost Card, it can cause considerable financial loss to the credit card holder. In the second kind of purchase, only little important information about cards (card number, expiration, payment, date, secure code) is required, this fraud is normally done on the internet or on Phone. To make a fraud, the fraudster has to know the details of the card. Most of the time, the real cardholder does not know that there is someone else using his card. The only way to find out this type of fraud is to analyze the spending pattern of every user to detect any discrepancy with card.

Hidden Markov model is probably the simplest and easiest model that can be used for model Sequential data, i.e. data samples that are dependent on one another [4]. An HMM is a double random process embedded with two different levels, one is hidden and open to all others. The Hidden Markov Model is a limited set of states, each of which is associated with a probability distribution. Transitions between states are controlled by a set of possibilities called transition probabilities. A result or observation can be generated in a particular situation, according to the related probability distribution. This is the only result, the state is not visible for an external supervisor and therefore the state is "hidden" to the outside; hence the name is Hidden Markov Model.A Hidden Markov Model (HMM) based credit card FDS, which does not require fraud signatures and yet is able to detect frauds by considering a cardholder's spending habit [5].

Naïve Bayes classifier is also a well-known Bayesian Network that is based on Bayes theorem of conditional probability and hence, is a classifier based on probability which considers Naïve i.e., strong independence assumption. It was formerly introduced with some other name, into the text retrieval community as a baseline technique for categorizing text because there was a problem of deciding in which category the document belongs to, with word frequencies as the feature.

The Naïve Bayes machine learning classifier tries to predict a class which is known as outcome class based on probabilities, and also conditional probabilities of how many times it occurred from the training data. This kind of learning is very efficient, fast and high in accuracy for real-world scenarios, and is known as supervised learning. Also, this is highly efficient because it estimates the parameters by using very small training data which is used for classification and is based upon word independence. Though Naïve Bayes is quite simple to implement and understand and uses strong assumptions. It gives pretty accurate results and also it has been proven over and over the time that Naïve Bayes works effectively in various areas related to machine learning.

In Naïve Bayes, the basic concept to computing the probabilities of various categories given a text is performed by using joint probabilities of categories and words.

II. Literature Review

1. [SushmitoGhosh, Douglas L. Reilly] "Credit Card Fraud Detection using Neural Network"
   This author explain about increase in fraudulent credit card usage has stressed the fraud management systems currently in use at banks and other institutions that process credit card transactions. In a rigidly controlled test on real world data from Mellon Bank's credit card portfolio, a neural network-based fraud detection system has been shown to provide substantial improvements in both accuracy and timeliness of fraud detection. The feasibility study demonstrated that due to its ability to detect fraudulent patterns on credit card accounts, it is possible to achieve a reduction of from 20% to 40% in total fraud losses, at significantly reduced caseload for human review. Software implementing this neural network approach to fraud detection has been successfully installed and integrated into the production environment on Mellon Bank's mainframe computer, and Mellon is achieving fraud loss reductions consistent with those predicted in the study.
2. [Vaishali] "Fraud Detection in Credit Card by Clustering Approach"
   This author explains about Clustering algorithm when implemented showed significant results. Most of the fraudulent activities could be correctly identified. However, there were quite a few non-fraudulent activities, which wrongly got detected as frauds. To detect the fraud accurately and efficiently, it is necessary that the real data should be available. It is found that the transaction is either fraud or legitimate through K-means clustering algorithm easily. The future work will be to improve the fraud transactions by using simulated annealing. For this the real data is necessary thing to improve credit card fraud
3. [John O. Awoyemi , Adebayo O. Adetunmbi, Samuel A. Oluwadare] "Credit card fraud detection using Machine Learning Techniques"
   This author explains about comparative performance of Naïve Bayes, K-nearest neighbour and Logistic regression models in binary classification of imbalanced credit card fraud data. The rationale for investigating these three techniques is due to less comparison they have attracted in past literature. However, a subsequent study to compare other single and ensemble techniques using our approach is underway. Performance of classifiers varies across different evaluation metrics. Results from the experiment shows that the KNN shows significant performance for all metrics evaluated except for accuracy in the 10:90 data distribution. This study shows the effect of hybrid sampling on the performance of binary classification of imbalanced data. Expected future areas of research could be in examining meta-classifiers and Metalearning approaches in handling highly imbalanced credit card fraud data. Also effects of other sampling approaches can be investigated.

4. [B SathishBabu, Rajeshwari] "Real-time credit card fraud detection using Streaming Analytics"
   This author explains about successful in reducing the rate of false alarms that can be achieved by comparing the transactions that were marked fraud with the transactions that were actual .The number of false predictions is reduced as we make use of enhanced techniques for finding the spending behavioral pattern of the cardholder. Streaming analytics helps in detecting and preventing the false transaction by making use of both offline data (History) and real-time data. Apache Spark will process the transaction data in real-time and the Hidden Markov model determines the fraud in the incoming transaction. Streaming analytics not only prevents fraud but also reduces the false alarm rate. The rate of false alarm is reduced by examining the relationship between the transactions that were recorded as actual frauds and the transactions which were guessed as fraudulent. We train the model according to the original card holder's data and it is frequently updated

III. Research Methodology

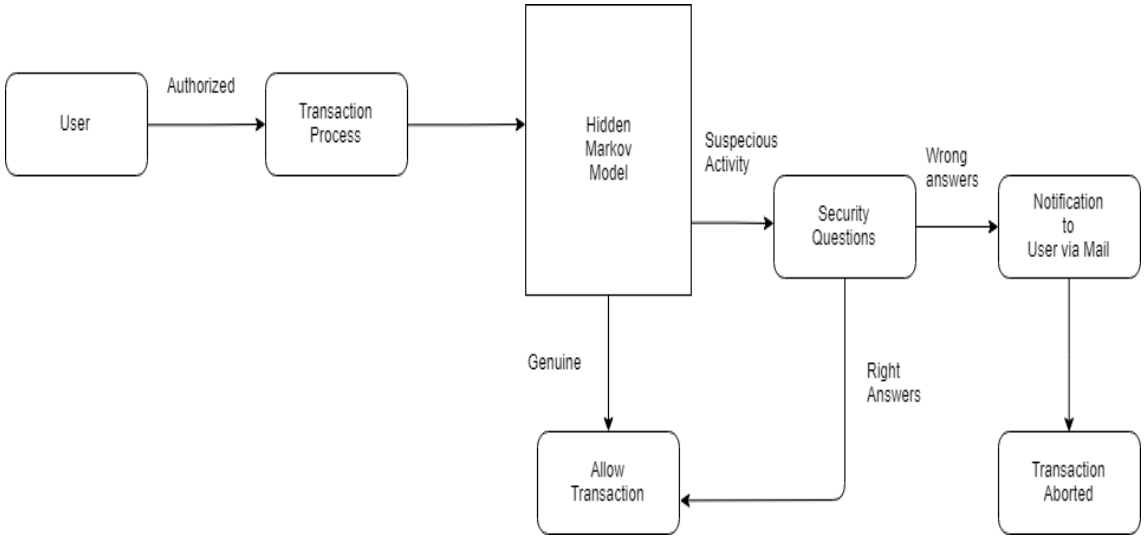

Figure 1: System Architecture

1. Data Analysis

The dataset which has been selected and used holds the records of European cardholders who made transactions using their credit cards in the month of September 2013. This dataset holds the record of transactions that were made within two days and total transactions made within two days are 284,807 transactions from which 492 transactions were found as fraudulent which makes the dataset highly imbalanced, more oriented as the positive class i.e., fraud transactions are 0.172% out of total transactions. And the dataset is in CSV format i.e., in a format where the data values are separated by commas. As PCA transformation of input values has been done in the dataset which makes the dataset contain only numerical input values. Unfortunately, the source of dataset did not provide the more background information. Original features and information due to confidentiality issues principal components that were obtained by PCA transformations are nothing but the numeric values under attributes V1 to V28 and the only feature that were not transformed using the transformation by Principal Component Analysis are the attributes or features 'TIME' and 'AMOUNT'.

'TIME' attributes or feature holds the data that denotes the elapsed time in between the first transaction of the dataset and each transaction. And the attribute or feature 'AMOUNT' hold the data which represents nothing but the amount of the transaction and this feature can also find its use for cost-sensitive and example-dependent machine learning. And finally the attribute or feature 'CLASS' is the response variable and it takes values '1' in the case of fraudulent transaction i.e., positive result and value '0' in the case of the genuine transaction.

2. Data Cleaning

Data cleaning is also called and known as Data cleansing because in this process the inaccurate and corrupted records from the dataset or a record-set or a table are identified and corrected i.e., removed and also this process focuses on identification of incorrect, irrelevant, inaccurate or incomplete parts of the data and then modification of that particular part by replacing it with some different value or completely deleting the dirty data.

In this dataset, the very first row wasn't required for the actual implementation of the algorithms which are used though it helped to analyze the dataset easily because that row is nothing but the attributes of the present data in the dataset and this is the reason why we are performing this step of data cleaning after understanding the dataset in the first step, so in this process of data cleaning or data cleansing, that particular row will be deleted from the dataset.

3. Implementing Hidden Markov Model

Hidden Markov model is probably the simplest and easiest model that can be used for model Sequential data, i.e. data samples that are dependent on one another.The implementation techniques of Hidden Markov Model in order to detect fraud transaction through credit cards, it create clusters of training set and identify the spending profile of cardholder [6].An HMM is a double random process embedded with two different levels, one is hidden and open to all others. The Hidden Markov Model is a limited set of states, each of which is associated with a probability distribution. Transitions between states are controlled by a set of

possibilities called transition probabilities. A result or observation can be generated in a particular situation, according to the related probability distribution. This is the only result, the state is not visible for an external supervisor and therefore the state is "hidden" to the outside; hence the name is Hidden Markov Model.

Figure 1: HMM Output

4. Implementing Naïve Bayes

The Naïve Bayes machine learning classifier tries to predict a class which is known as outcome class based on probabilities, and also conditional probabilities of its occurrence from the training data. This kind of learning is very efficient, fast and high in accuracy for real-world scenarios, and also this learning type is known as supervised learning.Bayesian networks ease many of the theoretical and computational difficulties of rule based systems by utilizing graphical features for representing and managing probabilistic knowledge [7].Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data [8].

Figure 3: Naïve Bayes Output

IV. Comparing the Results

	Naïve Bayes	**HMM**
Accuracy	0.8543	0.9581
Limitations	In order to achieve high accuracy. In classification tasks we need a big data set to make reliable estimations of the probability of each state.	Cannot express dependency between hidden states.

It is difficult to test the proposed method using real data. Banks do not, in general, agree to share their data with researchers. There is neither benchmark data set available for experiments. Therefore, we use a simulator to generate the transaction data that is also used in most of the related literature. We use the amount-driven simulator presented in. The number of normal transactions in a given length of mixed transactions is normally distributed with cardholder specified μ(mean) and δ(standard deviation).with this our Credit Card Fraud Detection has been done using Naïve Bayes and hidden Markov Model with the precision of approximately 88% and 95% respectively.

V. Conclusion

We have proposed an application of Hidden Markov Model for credit card fraud detection. The different steps in credit card transaction processing are represented as the underlying stochastic process of a Hidden Markov Model. We have used the ranges of transaction amount as the observation symbols, whereas the types of item have been considered to be states of the Hidden Markov Model. We have propose a method for finding the spending profile of cardholders, as well as the application of this knowledge in deciding the value of observation symbols and an initial estimate of the model parameters. It has also been explained how the Hidden Markov Model can detect whether an incoming transaction is fraudulent or not. Comparative studies reveal that the Accuracy of the system is close to 95% over a wide variation in the input data. The system is also scalable for handling large volumes of transactions.

References

[1]. HetaNaik, Credit Card Fraud Detection for Online Banking Transactions, InternationalJournal for Research in Applied Science & Engineering Technology (IJRASET**)** ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887Volume 6 Issue IV.

[2]. AbhinavSrivastava, AmlanKundu, ShamikSural, Arun K. Majumdar, Credit Card Fraud Detection Using Hidden Markov Model, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 5, NO. 1, JANUARY-MARCH 2008.

[3]. Shailesh S. Dhok, Credit Card Fraud Detection Using Hidden Markov Model, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[4]. V. Bhusari, S. Patil, Application of Hidden Markov Model in Credit Card Fraud Detection, International Journal of Distributed and Parallel Systems (IJDPS) Vol.2, No.6, November 2011

[5]. Divya Singh, Asst. Prof. RakeshPandit, Credit Card Fraud Detection Using Hidden Markov Model,International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January- 2015 1488 ISSN 2229-5518.

[6]. GauravMhatre, Credit Card Fraud Detection Using Hidden Markov Model, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2053-2055

[7]. Milgo, Carolyne, A Bayesian Classification Model for Fraud Detection over ATM Platforms, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. IV (Jul.-Aug. 2016), PP 26-32

[8]. R. Mallika1, Fraud Detection using Supervised Learning Algorithms, International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Issue 6, June 2017