March - April 2020
ISSN: 0193-4120 Page No. 2193 - 2199

Fraud Analytics System for Credit Card Using Machine Learning

Vishal Puri, Department of Information Technology,Sinhgad College of Engineering, Pune, vvpuri.scoe@sinhgad.edu
Dr. A. Ramesh Babu, Hindustan Institute of Technology and Science, Chennai. arbbabu67@gmail.com
Aniket Dhormare, Department of Information Technology, Sinhgad College of Engineering, Pune, aniketdhormare@gmail.com
Mahesh Dhondge, Department of Information Technology, Sinhgad College of Engineering, Pune mahesh.dhonge@gmail.com

Article Info
Volume 83
Page Number: 2193 - 2199
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Abstract:
This paper provides information on Credit Card Fraud Analysis with the help of some Machine Learning Techniques. The proposed system uses K-Means, Naïve-Bayes and SVM algorithm to detect fraud in credit card transaction. K-means is an algorithm used to solve clustering problem while Naïve-Bayes and SVM algorithm are used for detecting and classifying credit card transaction into fraudulent and non-fraudulent by matching the pattern of fraud as fraud usually has some past pattern associated with it. Result is compared with each other on the basis of fraudulent and not-fraudulent transaction, based on accuracy and shown in the visualization format.

Keywords: *Credit card fraud, Machine learning, Fraud detection.*

I. INTRODUCTION

In recent years, there is an increasing growth in online transactions and people rely on it for most of their requirements due to its ease of use and speed. Although the credit card gives convenience to the users, it can cause many hazardous frauds. When credit card details are stolen and misused, the cardholder follows up to the bank to enquire about the transaction. Fraud takes adaptive behavior so it requires sophisticated and long-established data analysis for detection. So this gives rise to a need for a fraud detection analysis system that can detect and prevent online transactions from frauds.

This paper is aimed to improve the security of online transactions by implementing a system for detection and prevention of fraudulent transactions with help of various types of algorithms and developing a secured system based on machine learning algorithms capable of analyzing databases to detect fraud. The main purpose of this paper is to use machine learning algorithms to develop the proposed system for the detection and prevention of fraudulent credit card activity. This system will improve the overall security of financial transactions. This paper will make use of line-chart or a histogram to display the ratio of fraudulent and legit transaction.

II. LITERATURE SURVEY

The paper Singh etc. all [1] in 2018, uses K-means clustering and Bayesian algorithm for detecting the Fraud. But the Main Challenge is about tracking user's behavior as user behavior as well as changes with time. Also, users profile keeps on changing with time. People's profile and behavior are generally not static. So, tracking users' behaviors is challenging.

Talking about Awoyemi etc. all [2] in 2017, focused on Naïve Bayes, KNN, logistic regression. Bahnsen etc. all [3] 2013, is the one that uses Bayesian algorithm which requires learning from data but leads to fitting data more than required. It increases the complexity of the model and uses an additional degree of freedom. Also, Bayesian network get affected by newer instances

In 2012 taking into account Xu etc. all [4] used SVM as the main method for detecting fraudulent activity but it still has challenges about the choice of the

kernel function. N. Malini etc. all [6] paper of 2017 uses KNN and Outlier Detection Technique for detection of fraudulent and avoiding false alarm. These are very useful in the prevention and detection of fraudulent transactions.

Fraud making people normally keep their behavior similar to normal users and keeps on finding new ways of making fraud without getting detected by adopting normal behavior to look it as legitimate. For tracking such people, detection techniques need to be sophisticated. Here, we need to add one more point is that individual behavior may be different with every different card he owns. So, it is necessary to analyze both users as well as card spending behaviors.

In 2018, Pumsirirat, etc. all use RBM with Auto Encode Algorithm and deep learning, which help deal with the fraud with new patterns. Zareapoor etc. all [5] of 2012 uses different techniques such as NN, Fuzzy SVM, KNN, AIS for intelligent detection of fraud. But it has some issues in the handling of categorical data. E. Caldeira etc. all [9] in 2014, using a neural network for detection.

But as this is a neural network there are some related input weight that requires modification of the whole network after every new pattern. Tripathi etc. all [10] of 2013 uses Bayesian learning which checks the pattern of that suspicious-looking transaction by comparing it. It has an issue of low processing power and expensiveness, though it is efficient. Some authors have suggested using HMM. It makes system scalable suitable for huge no. of transaction but has less accuracy near to 80%.

Please use automatic hyphenation and check your spelling. Additionally, be sure your sentences are complete and that there is continuity within your paragraphs. Check the numbering of your graphics (figures and tables) and make sure that all appropriate references are included.

TABLE I LITERATURE REVIEW

Authors citation	***Method***	***Challenges***
Singh etc. all[1] and 2015.	K-Means clustering and Bayesian algorithm	Hard to track user's behaviors, as a user changes their behaviors frequently.
Awoyemi etc. all[2] and 2017.	Naïve Bayes, KNN, Logistic Regression	Increase complexities in-depth learning
Bahnsen etc. all[3]and 2013.	Bayesian algorithm	Overfitting problem
Xu et. al[4] and 2012.	SVM	Choice of kernel function
Zareapoor etc. all[5] and 2012.	Uses NN, Fuzzy, SVM, KNN, AIS	Issues in Handling categorical data
Pumsirirat etc. all[7] and 2018.	Uses RBM, Auto Encoder Algorithm.	Less prediction accuracy.
E. Caldeira etc. all[9] and 2014	Uses a Neural network-based approach	As it uses a neural network, for every new pattern detected, it needs retraining of the whole network.
Tripathi etc. all[10] and 2013.	Bayesian learning, Dempster-Shafer theory	Highly expensive and low processing power.

III. FRAUD ACTIVITY IDENTIFICATION BY MACHINE LEARNING

A. Recognition of Fraud by analyzing the past behavior and history of the customers.

March - April 2020
ISSN: 0193-4120 Page No. 2193 - 2199

The purpose of this paper is to prevent fraud activities with help of data mining, identifying the patterns and relationship between fraudulent activities, pattern matching algorithm, calculation of different statistical parameters, analyzing customer's profile for detection in a real-time fashion. The Bayesian learning approach uses two parameters first one being customer past spending history and physical location for calculation of risk score which helps in analyzing genuineness of the transaction.

The customer spending pattern can be identified by the K-Mean clustering algorithm and by physical location, by comparing the pattern of the physical location of the customer. Customer spending pattern is recognized in 3 categories: 1. Low, 2. Medium, 3. High and further subcategories. but it become difficult to analyze and recognize the as user spending pattern changes with time. Rather Than just customer spending patterns, we require to also take card history into account. As single users owing multiple cards may have different behavior in case of each card.

B. An Optimized SVM Model for Detection of Fraudulent Online Credit Card Transaction

For Detecting fraudulent transaction activities, this paper proposes an optimized SVM model that uses non-linear Support Vector Machine, Radial Basis Function and grid algorithm for sparse transaction data and for deciding optimal combination of parameters. For optimal Support Vector Machine, a kernel function is decided which includes a parameter, category, Q. P. selection.

This SVM model is applied to the Fraud detection System. By doing data conversion, selection, and feature extraction, after finding an optimal SVM model and comparing it with the hybrid model, It is observed that it has higher performance than a hybrid one.

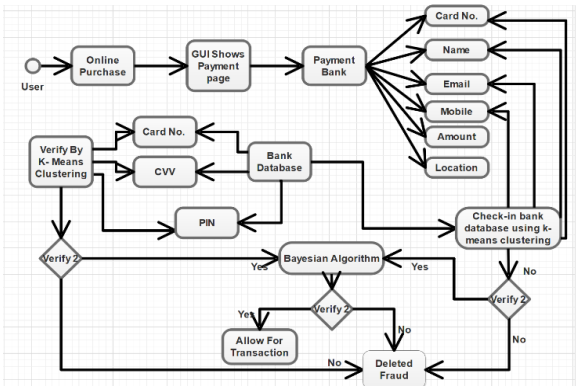

Fig. 1. Fraud analysis System

IV. METHODOLOGY

A. K-Mean Clustering algorithm:

K-Mean is used in solving cluster related problems. It uses Euclidean distance for the measurement of distances between instances. This is an unsupervised learning algorithm. Cluster is a collection of instances with similar properties. With the help of K-Means clustering, Similar instances are grouped into clusters. This is an iterative process where the instances get grouped into clusters with the highest available minimum Euclidian distance in each iteration.

K-means algorithm steps:

TABLE II. K-Mean Algorithm Steps:

Step 1:	Add libraries.
Step 2:	Make a data frame of the CSV file.
Step 3:	Create an array for the drop list and drop the elements from the data frame.
Step 4:	Apply k-means on the data frame.
Step 5:	Predict the labels for 'X'.
Step 6:	Print accuracy and confusion matrix.
Step 7:	Change the drop list and repeat from step 3 to step 6.

TABLE III. Pseudo Code:

Step 1:	Assign k means with randomly selected values

March - April 2020
ISSN: 0193-4120 Page No. 2193 - 2199

Step 2:	for a given number of iterations:
Step 3:	Iterate through items:
Step 4:	Find the mean closest to the item.
Step 5:	Assign item to mean.
Step 6:	Update means.

The objective function is:

$$J= \sum\sum w||x^i-\mu_k||^2$$

where,
μk= centroid of cluster, x^i = data point.

The Euclidian Distance(Eij) between (ai, aj) is given by

$$E_{ij}= sqrt(\sum(a_i - a_j)^2)$$

B. Bayesian Belief Network

Naïve Bayes for categorization of data by using probabilistic model with the help of some certain independent attribute for classification purpose by applying Some decision rules such that ML, MAP, recalibrated likelihood. This involves the estimation of a parameter in the distribution for training model.

TABLE IV. Naive-Bayes algorithm steps:

Step 1:	Add libraries.
Step 2:	Make a Data frame of the CSV file
Step 3:	Define split data function to drop the drop list items and to split the data into the train test.
Step 4:	Define get prediction function to fit the classifier to the dataset and to predict y values.
Step 5:	Take a drop list call split data function
Step 6:	Call get a prediction for the training dataset.
Step 7:	Go to step 5 and step 6 for the different drop lists.

C. Support Vector Machine

SVM Machine is used in prediction categorization into Legitimate Transaction and Fraudulent one. This has a supervised approach in the learning algorithm. It maximizes dist. (nearest data-point, hyperplane) and chooses the best hyperplane based on distance.

Margin = dist. (nearest training instances, decision boundary)/ ||weight factor||.

1) SVM Kernels:

SVM kernel is used to convert low dimension space into a higher one by adding dimension to it. Kernel trick is used to make linear model into non-linear one. The kernel is useful in transforming the problem into separable one, it makes SVM more flexible and correct.

TABLEV. Kernel Types

Linear kernel	$K(a, a_i) = \sum(a*a_i)$ where, a and $a_i$are two vectors.
Polynomial Kernel	Used for identifying non-linear input. $K(a, a_i)= 1+ \sum(a*ai)^x$where, x= degree of polynomial, needs manual specification.
Radial Basis Function kernel i.e. RBF kernel	$K(a,a_i)=\sum(a*a_i)=$ $K(a,a_i)=exp(-(\|\|a-a_i\|\|^2)/(2\sigma^2)$, Where, $(-\|\|a-a_i\|\|$, is a Euclidian distance, σ = free para. And$\gamma=1/(2\sigma^2)$ *Here,* $0\leq \gamma \leq 1$; And good value=0.1

TABLE VI. SVM Algorithm Steps:

Step 1:	Add libraries.
Step 2:	Make a Data frame of the CSV file
Step 3:	Select the SVM classifier and fit the dataset.
Step 4:	Do Model Evolution by prediction of 'y' values for the test dataset.
Step 5:	Plot confusion matrix.

March - April 2020
ISSN: 0193-4120 Page No. 2193 - 2199

2) Data Pre-Processing and selection :

In Data Preprocessing, the dataset attributes are converted into numerical data. In data processing data is first converted into a suitable form and then required data is selected. In this data detection, validation, correction of errors, handling missing data are some of the activities that are carried out.

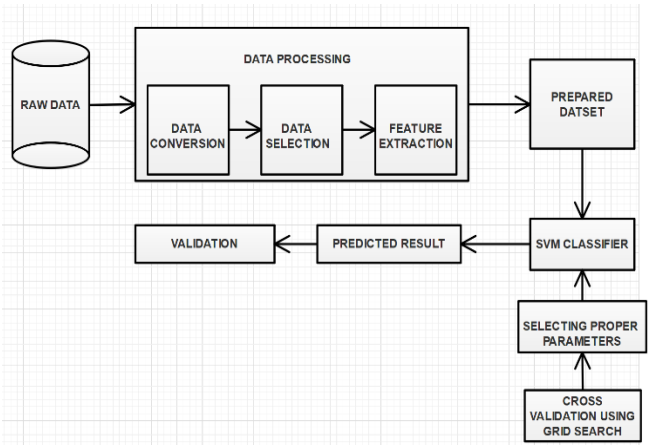

Fig 2. Data Pre-Processing and selection.

3) Feature Extraction:

Higher the number if feature higher is the difficulty and complexity in visualizing and working with training data set. So, reducing irrelevant and redundant features is important. This is done by dimensionality reduction. Which consists of two methods: 1) Feature Selection 2) Feature Extraction.

In Feature extraction, PCA i.e. Principal Component Analysis is used for reducing the set of features. PCA is the most popular method in dimensionality reduction for converting high dimensional data into low dimensional data by preserving required dimensions. It creates a new set of a data variable which can be used in place of original set of data variables.

4) *SVM Training And Classification:*

It is required to divide our dataset into training and testing datasets. So, we will divide it into 80:20 ratios.

Steps to perform:

Step 1:	Divide Dataset into legitimate and Fraud dataset
Step 2:	Ensure the random selection of data from each class by shuffling it.
Step 3:	Reduce the No. of features that are irrelevant in classification task

D. Result And Evaluation

Here is the Confusion Matrix for evaluation of false alarm rate and accuracy of detection.

TABLE VIII

	Confusion Matrix		
		Predicted	***Predicted***
Actual	Positive	TP	TN
Actual	Negative	FP	FN

In Confusion Matrix, column denotes predicted class and row denotes actual class. In Confusion Matrix,

True Positive: where Predicted output value is YES and Actual output value is also YES i.e. correctly predicted values. This shows the output value of transaction which are actually genuine and also got predicted as genuine.

False Positive: where Predicted output value is NO, but the Actual output value is YES. This shows incorrectly predicted genuine transactions as a fraudulent one. This leads to a false alarm situation.

True Negative: Where Predicted output value is YES but Actual value is NO. This shows an incorrect interpretation of fraudulent transactions as a Genuine one.

False Negative: Where Predicted is NO and Actual output value is also NO. This shows correctly detected Fake Transaction as Fraudulent one.

$$TP\ rate = \frac{TP}{TP + FN}$$

March - April 2020
ISSN: 0193-4120 Page No. 2193 - 2199

Diagonal of Confusion Matrix Gives Accuracy of the Matrix.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Error\ rate = \frac{FP + FN}{TP + FP + FN + TN}$$

V. Result of the Algorithm Used

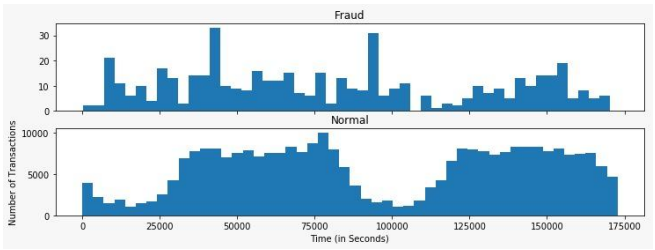

Fig 3. Subplot Visualization of fraud and normal transactions

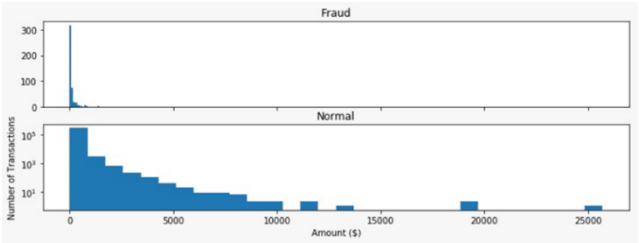

Fig 4. Subplot Visualization of fraud and normal transactions

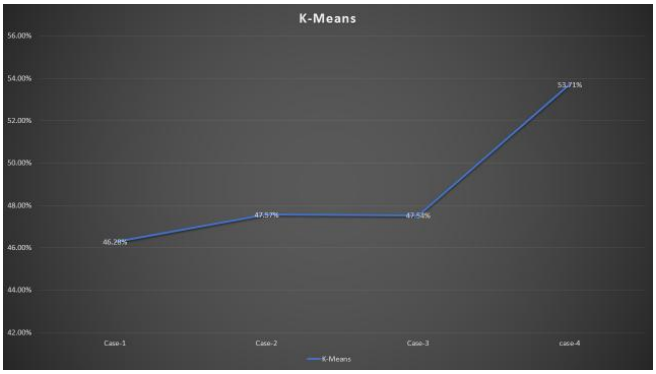

Fig 5. Visualization of K-mean

K means has less accuracy around 50% compared to that of SVM and Naïve-Bayes.

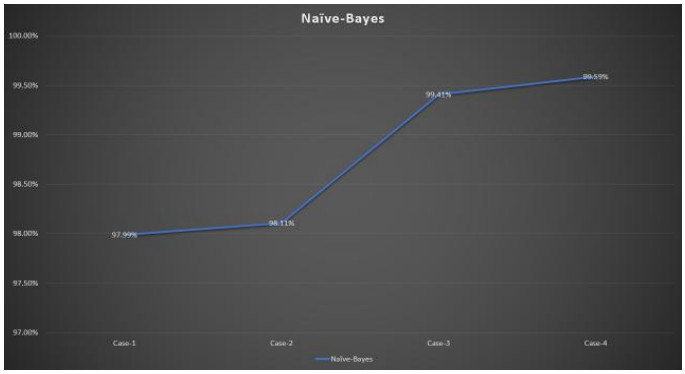

Fig 6. Visualization of Naïve-Bayes

Naïve-Bayes has accuracy around 98% and even goes beyond 99.5%.

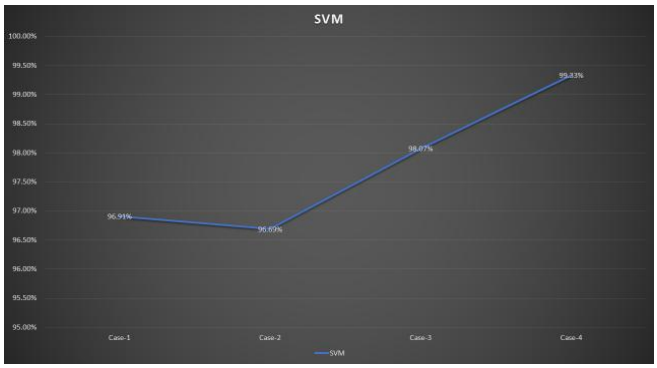

Fig 7. Visualization of SVM

SVM has an accuracy of around 99%.

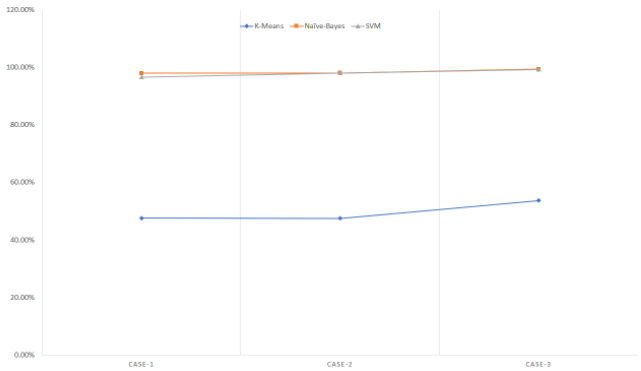

Fig 8. Visualization of K-means VS Naïve-Bayes VS SVM

March - April 2020
ISSN: 0193-4120 Page No. 2193 - 2199

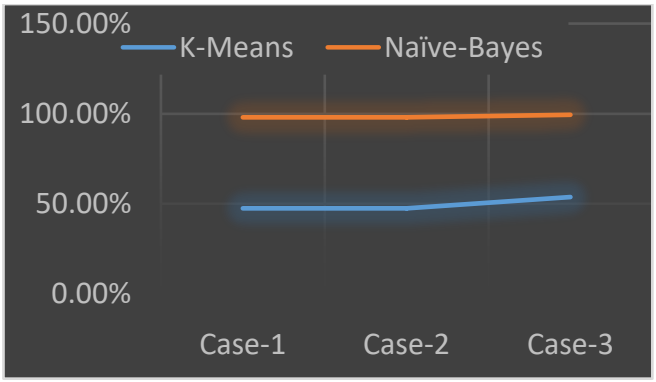

Fig 9. Visualization of K-means VS Naïve-Bayes

VI. Conclusion

The goal of a financial cybercrime detection model is to use machine learning to detect fraudulent activity and use various techniques based on data mining such as the K-means algorithm, Bayesian Learning Approach and Support Vector Machine and comparative study of these techniques. Our study shows a comparative analysis of all three classifiers which we used. Initially, we present an approach for every online transaction; the risk score depends on various factors such as user profile, amount of transaction. K-means algorithm will form the clusters of transaction gives the accuracy of nearly 50%. Bayesian Theorem is used for evaluation of risk score, using posterior probability and it gives thee accuracy up to 98%. Support Vector Machine (SVM) gives an accuracy of up to 99%.

VII. References

[1] Singh, Parvinder, and Mandeep Singh. "Fraud detection by monitoring customer behavior and activities." International Journal of Computer Applications 111, no. 11 (2015).

[2] Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." In 2017 International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9. IEEE, 2017

[3] Bahnsen, Alejandro Correa, AleksandarStojanovic, DjamilaAouada, and BjörnOttersten. "Cost sensitive credit card fraud detection using Bayes minimum risk." In 2013 12th international conference on machine learning and applications, vol. 1, pp. 333-338. IEEE, 2013.

[4] Xu, Wei, and Yuan Liu. "An Optimized SVM Model for Detection of Fraudulent Online Credit Card Transactions." In 2012 International Conference on Management of e-Commerce and e-Government, pp. 14-17. IEEE, 2012.

[5] M. Zareapoor, S. K. Seeja and M. AfsarAlam“ Analysis on Credit Card Fraud Detection Techniques: Based on Certain Design Criteria,” Int. J. Comput. Appl., vol 52, no. 3 pp. 35-42, 2012.

[6] N. Malini and Dr. M. Pushpa, “Analysis on credit card fraud identification techniques based on KNN and outlier detection” in 3rd

[7] International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17).

[8] ApapanPumsirirat, Liu Yan” Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine” in (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 9, No. 1, 2018

[9] Mr. Sunil S Mhamane, Mr. L.M.RJLobo.”Internet Banking Fraud Detection Using HMM.” In IEEE-20180, 2018.

[10]E. Caldeira, G. Brandao and A. C. M. Pereira, “Fraud Analysis and Prevention in e-Commerce Transactions”, 9th Latin American Web Congress, OuroPreto, pp. 42-49, 2014.

[11]K.K. Tripathi and R. Lata, “Hybrid Approach for Credit Card Fraud Detection”, International Journal of Soft Computing and Engineering, vol. 3, no. 4, pp. 8-11, 2013