IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 10, Issue 1, January 2020, ||Series -I|| PP 10-14

Analysis of Credit Card Fraud Data using PCA

V. Venu Madhav[1], K. Aruna Kumari[2]

[1]M.tech Student, Computer Science and Engineering, SRKR Engineering College, Bhimavaram, India
[2]Assistant Professor, Computer Science and Engineering, SRKR Engineering College, Bhimavaram, India
Corresponding Author: V. Venu Madhav

Received 24 December 2019; Accepted 08 January 2020

Abstract: Credit card fraud is a type of identity theft that pose a serious problem to financial services. Every year, fraudulent credit card transactions cause billions of dollars of loss. Detecting credit card fraud is a major problem among various researchers. Many researchers have proposed various approaches to credit card fraud detection, few among which use machine learning algorithms. The current work is based on the Principal Component Analysis (PCA) technique along with the K-Means clustering algorithm for the analysis of credit card fraud data. PCA is a tool in data analysis for dimension reduction with weighted voting. The proposed system exhibits the analysis results of the credit card fraud data in the form of clusters for better visualization. This system is suitable for online applications which have memory or computation limitation.

Keywords: Credit card fraud detection, Data Mining, K-Means and PCA.

I. INTRODUCTION

Credit cards are commonly used because of the growth of the internet business and the development of portable smart gadgets. Cardless transactions are presently ruling the internet world. For example, online transaction with a credit card is more popular, especially all credit card transactions are done by web payment gateways, e.g., PayPal and Alipay [1]. The credit card has made an online transaction easier and more convenient. The use of credit cards is widely increased, due to this fraud appears as major problem in credit card business. However, there is a growing trend of transaction fraud resulting in great losses of money every year. It is estimated that losses will be increased every year at double-digit rates by 2020 [2]. Since the credit card isn't required in the online transaction environment and the card's data is enough to finish payment, it is easier to conduct a fraud than before. Credit card fraud has become a top barrier to the development of e-commerce and has a dramatic impact on the economy. Fraud detection is a process of observing the transaction behavior of a cardholder so as to recognize whether an incoming transaction is done by the cardholder or others [3].There are two sorts of techniques for fraud detection: Misuse detection and Anomaly detection. Misuse detection uses some classification techniques to decide if an incoming transaction is a fraud or not [4]. Normally, such a method needs to think about the current types of frauds to make models by learning different fraud patterns. Anomaly detection is to construct the profile of normal transaction behavior of a cardholder dependent on his/her historical transactional data, and decide a recent transaction as potential fraud if it deviates from the normal transaction behavior. However, this technique requires enough successive sample data to characterize the cardholder's normal transaction behavior.

Anomaly detection is an issue of finding patterns in data that does not follow normal behavior [5]. These non-fulfilling patterns are frequently called as outliers or anomalies. Fraud detection is a system for detecting such anomalies or outliers in the information. Anomaly detection is used in a wide range of applications, for example, forgery recognition of Master cards, fault detection, intrusion detection, event detection in sensor networks and system health monitoring, etc. It is also used in pre-processing to remove noise and inconsistent data. Anomaly detection is an unsupervised data learning problem. It is observed that removing or adding abnormal data instance will affect the principal direction than removing or adding a normal one does. Using this Leave One Out (LOO) methodology, the principal direction of the dataset can be calculated without the existing target data instance. The outliers of data instance can be determined by the variation of resulting principal direction.

II. LITERATURE REVIEW

Mohammad et al., [6] explained the most popular types of credit card frauds that are used for detection. Basically, it implements that there are two types of credit card fraud i.e application fraud and behavior fraud. Almost all the existing work about the detection of credit card fraud is to capture the behavior patterns of the cardholder and to detect the fraud transactions based on these patterns. Based on previous studies of credit card frauds Srivastava et al., [7] focused on attacks mainly on behavior frauds, therefore this work used an approach called HMM for detection of frauds. An HMM is initially trained with the normal behavior of the cardholder. If the current transaction is not accepted by the trained HMM with a high probability it is considered to be fraudulent. HMM model can detect the only known type of fraud for which it has trained and it cannot detect new types of fraud.The main limitation of HMM is that it takes more time for training the data. To overcome this limitation Shiyang et al., [8] proposed a method for credit card fraud detection called the Random Forest approach. Random forest is used to train normal and fraud behavior features. Random forest is a classification algorithm based on the votes of all base classifiers. This method can detect fraud in real-time and it also trains the model in less time. However, this method can obtain good results on small data sets and cannot handle large data sets. Vaishali et al., [9] explained the credit card fraud detection using a clustering approach. Fraud detection methods are continuously developed to defend criminals in adapting to their strategies. Data is generated randomly for a credit card then the K-means clustering algorithm is used for detecting the transaction whether it is fraud or legitimate. Clusters are formed to detect fraud in credit card transaction which is low, high, risky and high risky. K-means clustering algorithm is a simple and efficient algorithm for credit card fraud detection.

Jyoti et al., [10] proposed data mining techniques for credit card fraud detection. Classification moels based on decision trees and visual cryptography are applied to the credit card fraud detection problem. Based on the historical data decision trees can be used to predict whether a transaction is fraudulent or normal one based on the probability. By using these classification models, financial losses due to fraudulent transactions can be decreased. Sahin et al., [11] used the decision tree and SVM for detecting credit card fraud. In this approach, the dataset is divided into three groups, which are different in the ratio between fraudulent transactions and legitimate ones. Here seven decision trees and SVM based models are developed for comparison. The experimental results reveal that the decision tree-based model is better than the SVM model. However, if the size of the training dataset is increased the accuracy of the SVM based model would reach the same performance which is equal to the decision tree.

III. IMPLEMENTATION

The proposed work is based on the PCA and K-Means for the analysis of credit card fraud data. PCA is used for extracting multi features at a time. First PCA is applied on the fraud dataset. Then PCA creates new features from original features. These new features are named as PCs. By making use of this PCs (new features) the relation between multiple attributes or associated attributes can be represented in the form of graphs for analysis purpose. The K-Means algorithm is applied on the PCs which is used to identify the similarity index of associated attributes.

1. Principal Component Analysis (PCA)

PCA is a method for dimensionality reduction, often used to reduce the dimensionality of large data sets, by converting a large number of variables into a smaller number of variables that still contains most of the information in the large set. PCA is used to extract multiple features at a time by applying some transformations.

Step 1: Standardize the data first.

The input data are standardized or normalized to allow each attribute to fall within the same range. This step helps to ensure that larger domain attributes do not dominate attributes with smaller domains.

Step 2: Calculate data point’s covariance matrix X.

The purpose of this step is to understand how the variables of the input data set variables to vary from the mean to each other, or in other words, to see if there is any relationship between them.

Step 3: Calculate the Eigenvectors and their related Eigenvalues.

PCA calculates k-orthogonal vectors (or) eigenvectors which provide basic standard input data. These are unit vectors, each of which points in a direction perpendicular to the others. These vectors are referred to as PCs. The input data is the linear combination of the PCs.

Step 4: Sort the eigenvectors in decreasing order according to their eigenvalues.

The PCs are arranged in the descending order based on its significance (or) strength. The PCs basically provide important information about variance and serve as a new axis for the data.

Step 5: Select the first k eigenvectors and that will be the new k dimensions.

Now select only the first k- PCs with the highest variance and remove the weaker PCs with the lowest variance.

2. K-Means Clustering Algorithm

K-means clustering is an algorithm that divides or clusters N data points into K sub-sets. K-means algorithm is the easiest and most common algorithm for clustering.

a) Select k objects from D randomly as initial cluster centers;
b) Repeat
c) (re)assigns each object to the most similar cluster, based on the average value of the objects in the cluster;
d) update the mean of the cluster, i.e., for each cluster calculate the mean value of the objects.
e) until there is no change.

IV. RESULTS

The credit card fraud dataset is taken as the input and the analysis results with respect to different attributes are shown in the form of graphs.

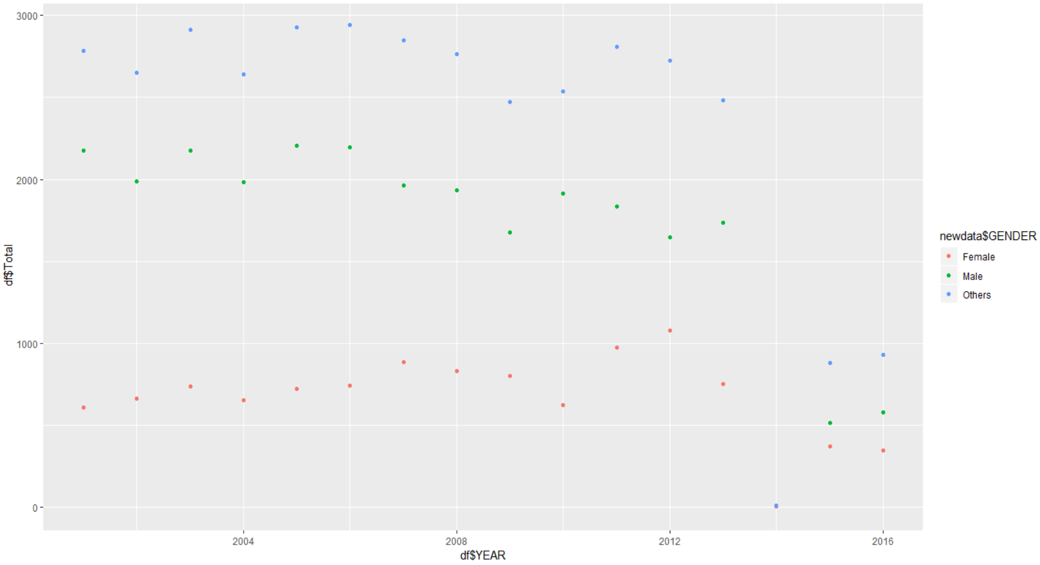

Figure 1: Analysis of fraud transactions w.r.to gender, year and total

Fig.1 explains the total number of fraud transactions done by males, females and others of Andhra Pradesh state from the years 2001 to 2016. Red color denotes the fraud transactions done by females, green color denotes the fraud transactions done by males and blue color denotes the fraud transactions by others.

Table 1: Optimized principal component values

	PC1	PC2	PC3
1	-3.1816	2.6012	1.8195
2	1.1808	2.3848	1.3465
3	-6.1650	2.4169	3.8636
4	0.0573	0.7617	-2.1505
5	2.2368	1.9841	-0.6204
6	-1.9089	0.3030	-2.135
7	-1.4725	1.4624	-0.4582
8	2.2214	1.7850	-0.6068
9	-3.4931	0.9311	-0.4875
10	-0.5515	0.5450	-0.9596

By using PCA new features are created by applying some transformations which are shown in Table 1. These features are called the optimized principal components.

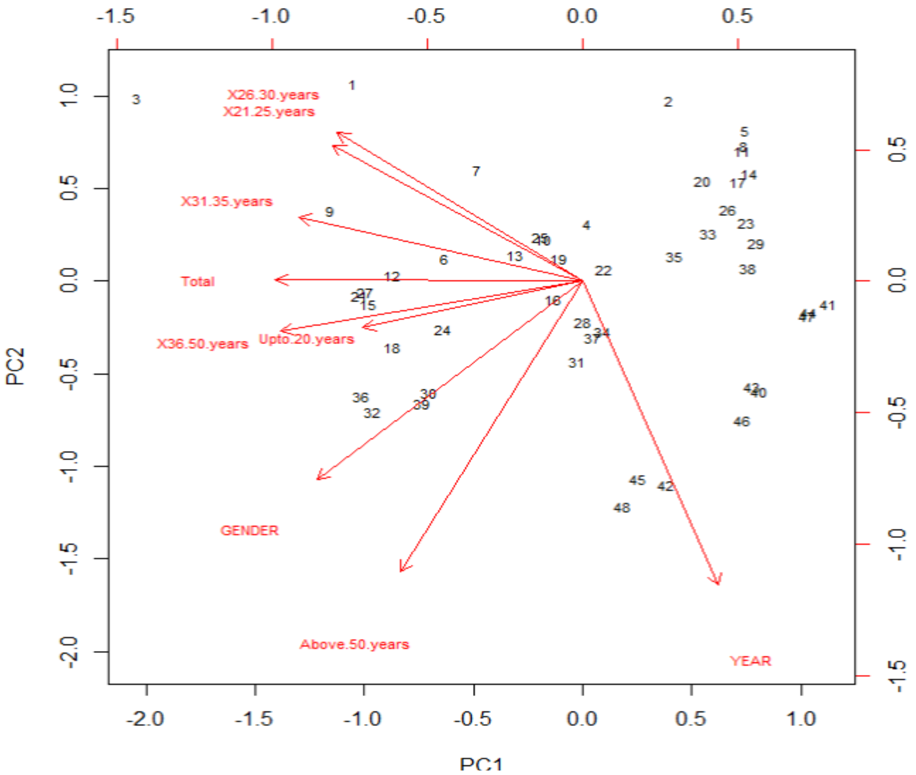

Figure 2: Bi-plot graph

Fig.2 is a Bi-plot graph which is used to visualize how the attribute values are related to one another. From new features PC1 and PC2 multiple attribute values are represented in a single dimension.

V. CONCLUSION

This work gives the technique for credit card fraud analysis depending on the PCA. PCA is used for extracting multiple features at a time from real-world datasets. Each PC represents an optimized value of the original attribute. In this approach, k-means clustering is also used for the identification of similarity index of associated attributes. Experimental results show efficient analysis of associated attribute relation. Furthermore, the proposed approach shows an optimized cluster representation effectively.

REFERENCES

[1]. Gupta, Shalini, and R. Johari, "A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant", International Conference on Communication Systems and Network Technologies IEEE, pp: 22-26, 2011.

[2]. Y. Gmbh and K. G. Co, "Global online payment methods: Full-year", Tech Report published at "YSTATS.COM", pp: 2-15, 2017.

[3]. E. Duman, M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search", Expert Systems with Applications, 38(10), pp: 13057-13063, 2011.

[4]. Ju, W. H, and Vardi, Y, "A hybrid high-order Markov chain model for computer intrusion detection", Journal of Computational and Graphical Statistics, 10(2), pp: 277-295, 2001.

[5]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey", ACM Computing Surveys, vol. 41, no. 3, pp: 15:1-15:58, 2009.

[6]. Mohammad, B., Barone, L., Bennamoun, M., and French, T., "Nature-inspired techniques in the context of fraud detection", IEEE Transactions on Systems Man and Cybernetics-Part C, 42(6), pp: 1273-1290, 2012.

[7]. Srivastava, A., Kundu, A., Sural, S., and Majumdar, A., "Credit card fraud detection using Hidden Markov Model", IEEE Transactions on Dependable and Secure Computing, 5(1), pp: 37-48, 2008.

[8]. Shiyang, X., Guanjun, L., Zhenchuan, L., Lutao, Z., Shuo, W., Changjun, J. "Random Forest for Credit Card Fraud Detection", IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), pp: 1-6, March 2018.

[9]. Vaishali, "Fraud detection in Credit Card by Clustering approach", International Journal of Computer Applications, vol. 98, no.3, pp: 29-32, 2014.

[10]. Jyoti, R. Gaikwad., Amruta B. Deshmane., Harshada V. Somavanshi, Snehal V. Patil, Rinku A. Badgujar, "Credit Card Fraud Detection using Decision Tree Induction Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 4, no.6, pp: 66-69, November 2014.

[11]. Sahin, Y., and Duman, E., "Detecting credit card fraud by decision trees and support vector machines", International MultiConference of Engineers and Computer Scientists (IMECS), vol.1, pp: 442-447, 2011.

V. Venu Madhav1.et.al. "Analysis of Credit Card Fraud Data using PCA." *IOSR Journal of Engineering (IOSRJEN),* 10(1), 2020, pp. 10-14.