International Journal of Security and Its Applications
Vol. 12, No. 5 (2018), pp.21-36
http://dx.doi.org/10.14257/ijsia.2018.12.5.03

Analyzing the performance of Various Fraud Detection Techniques

Anamika G[1], Mayuri K[1], B. Kharthik Kumar Reddy[2], N.Ch.S.N.Iyengar[2] and Ronnie D. Caytiles[3]

[1]*SCOPE, VIT, Deemed University, Vellore*
[2]*Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, India*
[3]*Multimedia Engineering Department, Hannam University, Daejeon, Korea*
srimannarayanach@sreenidhi.edu.in

Abstract

Sales of a product plays a vital role, towards profitability of any organisation. It helps in building trust and allegiance between customer and business. Sales department in any organisation help us to build a strategic decision for growing their business. In the process of building a strong relationship, there is chance of performing fraud, termed as sales fraud. Detecting such fraud can be of great help to the organization and taking appropriate steps to prevent the same in future. In this paper, we have implemented various clustering techniques like, k-means, k-modes, Hierarchical clustering, partitioning around medoids and also the Self organizing map technique in order to efficiently analyse and thus detect the sales fraud. The finding in our research, states that the self organizing maps and Hierarchical clustering provides good classification accuracy. These algorithms are strongly recommended in the field of fraud detection.

Keywords: *k-modes, Self organizing maps, Hierarchical clustering, medoids, sales fraud*

1. Introduction

During sales, many organisation takes a feedback from the customer as this is very valuable for improving their products. Their responsibility to give pricing of product as per the customer satisfaction so that their purchase will increase in a great extent. They have to keep record of all pricing details of goods, product information and distribution of goods. Sales fraud occur when there is unwanted manipulation of records related to goods. Others include giving false information of sales of goods. When there is hiding of details of profit or loss and money spent in a particular period on goods is also a sign of fraud. Fraud can be done in purchase department, sales department and account department, all the department are related and have to maintain a ledger for their record which helps the company for analysing the profit or loss.

Sales fraud can be performed in various forms, physically stealing the goods or by giving false information in ledger. In both the cases, records are manipulated and it affects the organisation while taking a next strategic decision for their growth. If the fraud occurs through an employee or non-employee, it can be easily reflected back to the record, that depends on the type of goods being stolen. Broadly, sales fraud can be categorized in two forms, False Sale information and False Purchase information. In a sales transaction record, the records of both are maintained, selling of goods and money that is coming in. Having only the record of distributing goods for a long time, not having any information of money coming in, it may also reflect the false information of selling of goods. This

Received (June 14, 2018), Review Result (August 5, 2018), Accepted (August 19, 2018)

ISSN: 1738-9976 IJSIA

requires the explanation regarding the sales of the goods. whereas, False Purchase Information is opposite of false sales information in which supplier get paid by the amount but the delivery of the product is not satisfactory with respect to the amount paid. This can be recorded in inventory department as the amount is already deducted but the product details are not available. This type of false information can be generated by supplier or by employee of that organisation. The main aim of this fraud is to hide the details of sales to being recorded. For example, if a product is being sold to a customer and he is paying by cash and there is no receipt generated, then the company records show the loss of product details, actually it has been delivered to the customer. False returns is a fraud related to the returning of the money to the customer. Some employees are having authority to return the product under some circumstances but returning of the money is not in their policy. Sometimes money is returned to the customer, by deduction of excessive amount of refund to be done. Many businesses are having selling of records by reading a bar code. If the product is not having initialised bar code, then that record cannot be mentioned in the inventory which shows loss of product data. In this fraud, the salesperson is not involved, this can be directly done by organisation for hiding the product details so that they can get a profit without showing it to the government. Fraud risk analysis is the process of determining the probability of the occurrence of fraud, the way used to prevent it and the techniques used for detection of fraud. Various action plans can be planned which should be cost effective to the organisation. It is finding holes in organisation and filling up with an effective way. There are four types of actions can be implemented for analysing the fraud are as follows: **Identify the threats:** There are four main possible threats involved in any organisation, that are financial, informational, operational and strategic threats. Financial threats are the main source of fraud occurrence, loss of money, loss of financial information, money spent in unnecessary activities. **Estimating the loss and risk from threats**: In this step, we are estimating the probability of risk that can occur, reoccurrence of the loss of data, and what are the losses to be involved. **Identify the prevention steps**: After identifying the loss and estimating the risk, the businessman has to take proper decisions to prevent the occurrence of the threats and the way to deduct the fraud that can occur. Prevention of the fraud occurrence is better than deduction of fraud. **Analysing the cost evolved:** The businessman has to determine the cost that that is worthful for controlling the fraud that can occur. Before taking any step for controlling the fraud, they have to analyse effectively, thus preventing the organisation from the occurrence of fraud. No thumb rule is there for analysing the risk evolved in organisation. We have to take steps according to risk levels of fraud and which prevents the organisation and cost involved should be beneficial.

2. Literature Review

In [1], the author stated that online shopping is the big concern of fraud in the online transaction. In which, there is a need for retailer for checking the genuinely of customer. Towards resolving the problem addressed. The author, in [2] had extracted the data from transaction, which is usually carried out by aggregating the transaction to observe and thus analyse the spending behaviour of the customers. In this research, the scholars use the Von Mises distribution to create a different set of features by analysing the periodic behaviour of the customers. In understanding different types of fraud detecting techniques. The author, in [3] stated that fraud detection is not only detection of fraud but it includes the detection of activity that are likely to be fraud. Fraud is an activity growing as the business increases millions times. Various techniques, deep learning, artificial neural network, data mining, machine learning, fuzzy logic, genetic algorithm for detecting the fraudulent transactions. This paper describing the data mining algorithm in a detailed way by

calculating the confidence, cost factor and weights in neural network for getting high accuracy [27]. It also showing that hidden Markov method is having low accuracy. Also, discussed 13 different model with strength and weakness of model showing that performance of fraud detection is depends on train data. The future gap, In the work states that the accuracy of model can be improved by estimating the cost factor. In [4], the author used the big data technologies in order to detect fraud for the considered dataset. In which, scholars focused on processing large amounts of data in order to detect fraud in real time. In [5], the author presented a fraud detection of credit card by using an artificial neural network as well as the meta cost Method and observed that there is an increased rate of fraud detection and less cost needed in doing so. The method first trains the dataset using the neural networks. Then they try and re-label each of the records. This is done by passing each record through a number of neural networks and then averaging the result to find the probability of getting a genuine transaction or a fraud transaction and thus each record is relabelled. These relabelled records are given as an input to the artificial neural network to detect whether a given transaction is fraudulent or not. In [6], the author used the Predictive Analytics Technologies (PAT), and evaluated their capabilities, relevant criteria and features to prepare a score card. The different PAT vendors taken into consideration are Falcon Fraud Manager, IBM SPSS Manager, SAS Fraud Manager, Cyber Source Decision Manager, ACI Proactive Risk Manager, which are evaluated based on the relevant features and criteria and thus is the scorecard prepared. The hidden Markov model to detect credit card fraud that obtains a high fraud coverage and low false alarm rate [7]. Salazar et. al. (2016) [13], proposed a framework to address the problem of fraud detection of credit cards using Pattern recognition and Signal processing, wherein the number of transactions that are fraudulent are very less as compared to the legitimate ones. Likelihood ratio scores are fused together in order to provide a better solution for the problem of fraud detection. Surrogates are created from the real data, which is then pre-processed, then given as an input to the various classifiers where they are trained and tested to produce scores. These scores are then fused to estimate the KPI (Key Performance Indicators), which are then used to get the result. The detection methods used are LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), NGM, alpha-INT, COP, RF (Random forest), SVM (Support Vector Machine). Various set of features are used on the real credit card dataset of a European Credit card company to check their impact. The use of the proposed periodic features, the results showed an average increase in savings of 13%. A fraud detection system that overcomes the limitations of the existing fraud detection systems [11], like, the scalability issues, extreme imbalanced class and time constraints. This they do by using a hybrid support vector machine (HSVM), which is the most used method for pattern recognition and classification along with communal and spike detection. The raw data extracted is converted into required form, then given as an input to the HSVM classifier. The HMM in its implementation creates clusters which depict the spending profile of the cardholder. The transactions are clustered into three clusters based on whether the amount spent is low, medium or high. An artificial neural network technique is used at the payment gateway to check if the transactions are fraudulent or not, by using the data provided by the merchant and that present in the payment gateway [17]. Data, thus extracted is normalized to generate rules which are in turn provided to the classifier. The transactions are divided into four categories viz., fraudulent, doubtful, suspicious and not fraudulent. They also try and combine various classifiers in order to improve the accuracy of the prediction, whether the transaction is fraudulent or not. They involve using four layers, the first being the data storage layer, the second being the batch processing, the third sharing of key values and the last denoting the

detection of streamlines. Use of the latest big data technologies like Hadoop, HBase, sparkis used. Credit card fraud detection is becoming necessary as the use of credit card is increasing in day to day life [11]. Credit card is becoming advantageous as it reducing to carry cash in hand. Various algorithm is applied like clustering algorithm, prediction algorithm, forecasting algorithm, it can be used as both pre-processing of data as well as for credit card transaction. It will also try to reduce the error rate by reducing the rejection of fraud data. we can detect fraud as data mining model is giving low probability values. Frauds can be arising if it is lost or stolen or misused of the card unknowingly to user. This paper is completely based on the card transaction of the user for analysing the frauds, by reducing false positive transactions. Feed forward neural network is implemented and we can extend the work by combining back propagation network. A new KDA model, namely k-means clustering, DBSCAN, agglomerative clustering algorithm using RapidMinertool [15], proposed a KDA model yielding 81% result from dynamic data and 68% from historical data, which helps to reducing the processing time. A clustering method for detecting a fraudulent transaction dataset used is generated randomly [10]. K-means algorithm is most basic algorithm of clustering and there is no method which will give you 100% accuracy but we can reduce the error rate of the approach. In this paper, we are categorising the fraud depends on the risk level. As the dataset used is randomly generated data, some of the frauds are misclassified and in similar way. For knowing the efficient working of model, we need to apply clustering approach in real data. We can also improve our work by applying simulated annealing. In [12], the presented different types of credit card fraud in the business industry and how it effects to user, merchants and banks. Fraud detection is not only to detect fraud but as quickly as possible we have to detect the fraud. They were explaining the loss due to fraud like cost. Their main focus on MasterCard and VISA transactions. As the online transactions increases, criminal activity is also increased as it is difficult to grab the actual offender. But as the technology increases we can able to reduce the fraud with respect of time. Machine learning algorithm can help the system for analysing the insights of the data for analysing the activity of customer. This paper combining the machine learning algorithm and neural network for highly detecting the fraud and reducing false alarm rate. Deep learning is used for analysing the behaviour of user from various transactions. Many industries using time series model by performing seasonal behaviour as per demand of product [8]. This paper discusses Holt Winters exponential smoothing for analysing seasonal time series which includes multiplicative seasonal model and additive seasonal model. Seasonality and trend can be exhibit by Holt Winters model. Mean Absolute Percentage Error is used for measuring error. L and Look back size are important parameter for analysing the performance of Holt Winters adaptive model. In author in [16], proposed a model combining ARIMA model and ANN for better performance as ARIMA is a traditional linear time series model whereas ANN is a latest non-linear model. By combining the feature of non-linear model and linear model so that we can strengthen our forecasting results. Hybrid model is having low variance and low error. As the changes in the data is unstable, so it is convenient to use hybrid model for reducing uncertainty of data. Spreadsheet is used for optimizing forecasting result of Holt Winters exponential smoothing [14]. This procedure is suitable for small size dataset like sales dataset. Spreadsheet can able to work in parallel with measures of forecasting error and calculating the initial values of the model component. They are getting local minima with RMSE which help us in choosing the parameters among forecast error and initial values. The author in [9], used the multiplicative seasonality for forecasting error or variance depends on trends and seasonal. Various model is provides updating of same error variance though it is necessary to choose most appropriate model and

simulation study is to find the difference between percentage of coverage. Maximum likelihood and correlation method is used for identifying the appropriate model. In [18] the author used weighted fuzzy logic to assign weights in training the data to extract sentiments from the labeled tweets and achieved good F-score. where as in [19] the author made a detailed comparison on predictive models and perform analysis on Time series dataset. In [20] the author perform analysis on PIMA diabetes dataset and predicted the levels of diabetes based on insulin feature. where as in [21] the author used gradient ascent algorithm in finding out the exact weights of the terms used in determining the sentiment of tweet and used Boosting approach to improve the accuracy of linear classifier. In [22], the author provides a novel way of performing prediction on Breast cancer dataset, compared the performance of three different feature selection algorithm and proved that genetic algorithm is giving best result in selecting the best feature among all the available feature. SVM algorithms gives the best result in predicting the level of certainty in breast cancer. In [23], the author made an attempt to develop a recommender system, helping in searching the item, that might out found by themselves, in which precision and recall measures are used in measuring the performance of proposed model. In [24], the author made a research in solving the problem in Diabetic Retinopathy. In which, the author proposed a Model, which can capable of calculating the weights, that gives severity level of the patient's eye by using weighted Fuzzy C-means algorithm. In [25], the author proposed a build a model for airlines, that can perform sentiment analysis on customer feedback and achieved Vital accuracy. Where as in [26], the author experimented on finding out the impact of feature selection on overall sentiment analysis and stated that Term frequency have greater impact on analyzing sentiments rather than bigram approach.

3. Problem Formulation

Sales frauds are one of the major problems faced by every organisation in world. The losses due to frauds are increasing rapidly from millions to billions. It is a challenge to build cost effective, risk analysing tools which can identify and avoid the fraud activities. The fraudulent activity can arise from the supplier, employees, non-employees and sales person. The cost incurred due to fraud detection involves the association fee, organisation loss, and administration cost, which must be should be minimised by finding an effective method of determining the fraud. As the times are changing and business are growing, the fraudsters are also developing their techniques to offend the activities. So, our model should also be able to change its behaviour according to the fraudster's activities. Different machine learning algorithm and data mining techniques are used for effectively determine the fraudulent as well as genuine transaction. It is very difficult to get real data, as the labelled data will not provide the deterministic result. Many financial institutions don't share the data because of the competitive growing world. We cannot make any comparison with real data and labelled data result. We are performing the various clustering approach and time series model for detecting the fraud in an efficient way. Clustering techniques helps us to determine outlier in fraudulent and in genuine transaction. Time series model that help us for predicting the values.

4. Methodology

The fraud dataset used is pre-processed to suit the techniques and algorithms to be applied. Here, applied the Clustering and Time series techniques in order to detect the fraud transactions.

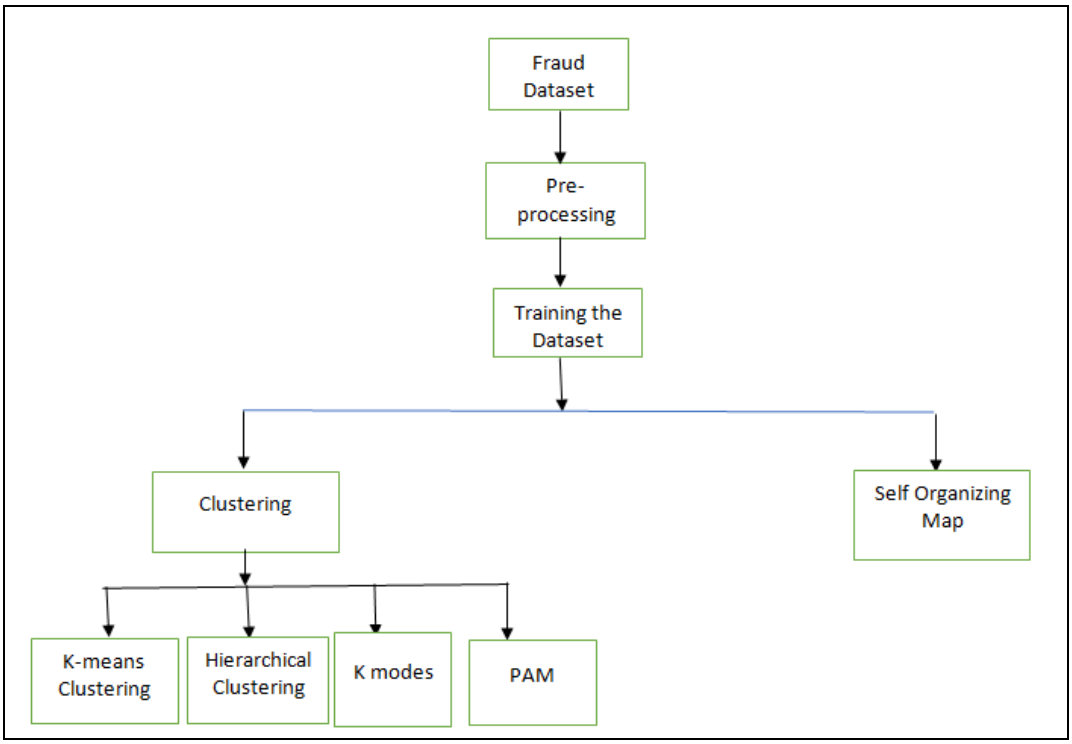


Figure 1. Methodology

4.1. Clustering Techniques

4.1.1. K-means Clustering: Clustering is a technique to divide the data into different groups. In which, the intra cluster similarity of the data must be high and the inter class similarity must be less i.e. the data points of different clusters should be dissimilar. k-means clustering algorithm is a widely used unsupervised algorithm. Given a set of *n* data points the goal is to cluster these *n* data points into *k* clusters, and the position of the clusters should be such that the distance of the data points from the cluster is minimum. The distance of the data point and the cluster centroid is calculated as follows:

$$d(cluster\,centroid) = \sqrt{(x - x_c)^2 + (y - y_c)^2}$$

Algorithm:

Step 1: Initialize the k cluster centroids representing the number of clusters in consideration.

Step 2: Calculate the distance of each point from the centroids initialized and place the data point in the cluster such that the distance of the data point from the centroid of the cluster is minimum.

Step 3: Calculate the new centroids of the clusters formed.

Step 4: Repeat steps 2 and 3 until the centroids don't change i.e. they converge.

4.1.2. K-modes Clustering:

K-modes clustering technique has been widely used in situations that involve the categorical data. The statistical inference measure such as mean cannot be calculated for categorical data, and k modes exactly handles this limitation of the k

means algorithm. A dissimilarity measure like hamming distance can be used for the data involving categorical data.

$$d(x, y) = \sum_{i=1}^{n} (Z_i, Q_i)$$

Algorithm:

Step 1: Initialize k clusters by choosing k initial modes that may or may not be present in the dataset in consideration.

Step 2: Calculate the similarity of the object with respect to the cluster modes available by the dissimilarity measure as:

If(object!=cluster_mode) then dissimilarity=1

If(object==cluster_mode) then dissimilariy=1-nrj/total

Step 3: Place the data point considered into that cluster for which the dissimilarity value is minimum.

Step 4: Recalculate the cluster modes based on the new cluster objects added.

Step 5: Repeat Steps 2 to 4 until the cluster modes remain the same *i.e.*, they converge.

4.1.3. Hierarchical Clustering:

Hierarchical Clustering is a clustering technique wherein the goal is to divide the given data points into a hierarchy of clusters. This clustering can either be done by the 'Agglomerative method of hierarchical clustering' or the 'divisive method of hierarchical clustering'. The Agglomerative method of hierarchical clustering, the bottom up approach, involves placing each data point or the data object into a cluster of its own. Then, based on the similarity of the objects, their respective clusters are merged to form a bigger cluster. This is repeated till all the similar objects are placed in a single cluster. The divisive method of hierarchical clustering, the top down method, involves placing all the data points or the data objects into a single cluster. Then, based on the dissimilarity of the data points they are put into different clusters. This process is repeated until all the data points that are dissimilar are placed into different clusters. Given a set of k clusters, and a *KXK* matrix of their distances.

Algorithm:

Step 1: Place all the data points into their own respective clusters. The distance between the clusters is equal to the distances between the data points they contain.

Step 2: Find a set of two clusters such that the distance between them is minimum and put them into the same cluster by merging the two respective clusters.

Step 3: Calculate the distance between the new cluster and the old cluster.

Step 4: Repeat steps 2 and 3 till all the data points are there in one single cluster. Different Hierarchical clustering techniques involve:

a) **Single linkage**: Represents the distance that is shortest between a point in one cluster and the data point in other cluster.

b) **Complete Linkage**: Represents the distance that is largest between the data point in one cluster to that of the data point in other cluster.

c) **Average Linkage**: Represents the distance that isthe average of the distances of the data point in one cluster and all the data points in other cluster.

d) **Centroid**: Represents the distance of the datapoints that denote the centroids of the clusters taken into consideration.

e) **Ward**: Represents the sum of squares of the data points of the two clusters.

4.1.4. Partitioning Around Medoids:

The k-means algorithm is sensitive to the outliers. A better solution is provided by the portioning around medoids (PAM) method of clustering. Here, the clusters are not represented by the centroids but the medoids. It supports all datatypes, continuous too.

Algorithm:

Step 1: First, randomly choose k medoids to represent k clusters.

Step 2: Calculate the distance of the data points from the cluster medoids.

Step 3: Assign the data point to that cluster that is closest to the medoid of the cluster under consideration.

Step 4: Add the distances of the data points from the medoids to get the total distance.

Step 5: Select a point that is not a medoid and swap it with the current medoid.

Step 6: Reassign every data point to the closest medoid's cluster.

Step 7: Calculate the total cost.

Step 8: If the calculated total is less then keep the new point as the next medoid.

Step 9: Repeat Steps 5 to 8 till the medoids converge.

4.1.4. Self Organizing Maps:

A competitive learning method, Self Organizing Maps (SOM) is the most used neural network technique. SOM is an unsupervised learning technique that does not need the intervention of the humans to learn and has no knowledge of the input data properties. The self organizing maps do not need the labels for the input data. It provides a mapping (that preserves the topology) from higher dimensions to the map units. The points that are nearer to each other are placed as adjacent map units. The nearby map units form a lattice and thus mapping is from a high dimension to a plane. plane. The SOM preserves the distance between the data points under consideration. Generalization can be done by the SOMs and thus can be used for analysis of the clusters. The self organizing maps create a feature map from the continuous space putting them into a discrete space.

Algorithm:

Step 1: Initialize the map by choosing random values for the initial weights.

Step 2: Perform sampling by choosing a set of vectors from the input space.

Step 3: Choose the neuron who weight is closest to that of the chosen vector.

Step 4: Apply the equation of updation

Step 5: Repeat steps 2 to 4 until the SOM becomes constant or does not change.

5. Implementation and Results

The proposed work is implemented in R 3.3.2. Dataset used is 'sales fraud detection' which consists of five attributes namely, ID, Prod, Val, Quant, Insp. The attributes are described as follows:

a) ID: Represents the sales person identification

b) Prod: Represents the identification of the product.

c) Val: Represents the reported transaction value.

d) Quant: represents the quantity of a particular product

e) Insp: Contains the report of three possible values i.e. ok, fraud and unkn.

Here, the dataset was in categorical form. We have pre-processed and converted the categories namely ok, fraud and unkn, so that the data can be effectively applied to various classifiers to be used. Ok is encoded as '0' which indicates no fraud, unkn is encoded as '1' which indicates that the status as to whether the transaction is fraud or not is unknown.fraud is encoded as '2' which means that the transaction is a fraud detection.

5.1. Clustering Techniques:

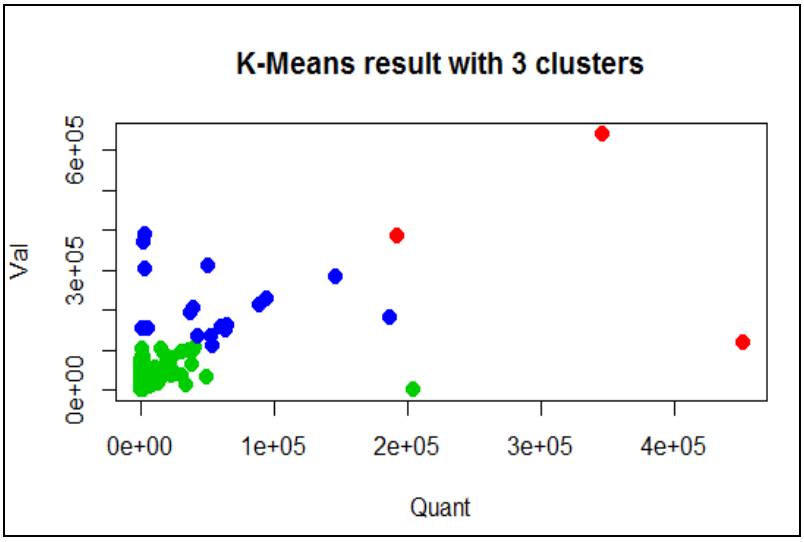


Figure 2. Performance of K-Means

K-means Clustering: We first consider the parameter k as three in Figure 2. The result we got is follows: The green dots represent the transactions that are not fraud, the blue dots represent the transactions for whom the status as to whether the transactions are fraud or not is unknown. The red dots represent the fraud transactions. The accuracy achieved for the above method is: 70%. We then apply the elbow method of finding the optimal number of clusters for the given scenario. The method used shows that the graph is constant after eight clusters thus we use eight clusters for our dataset as shown in Figure 3.

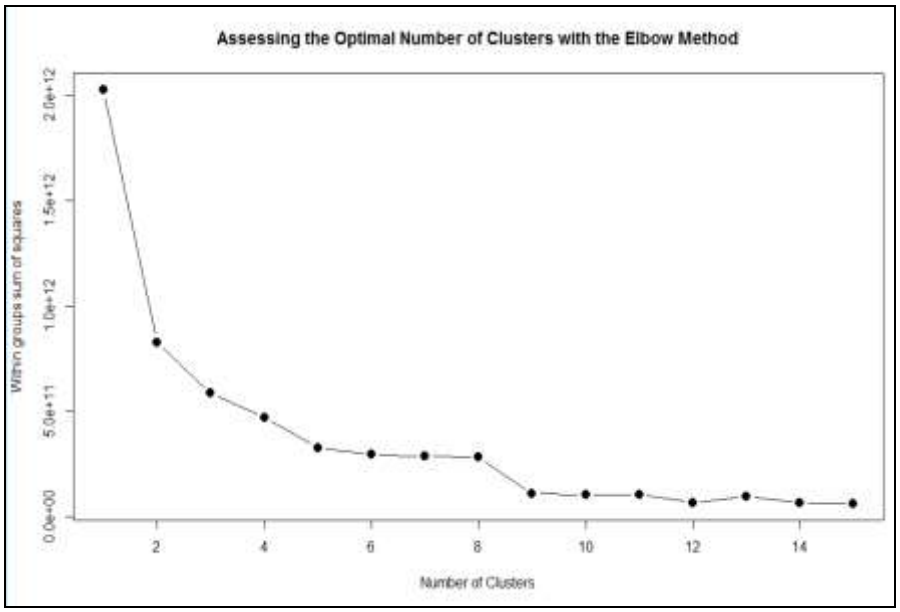


Figure 3. k-means when k=8

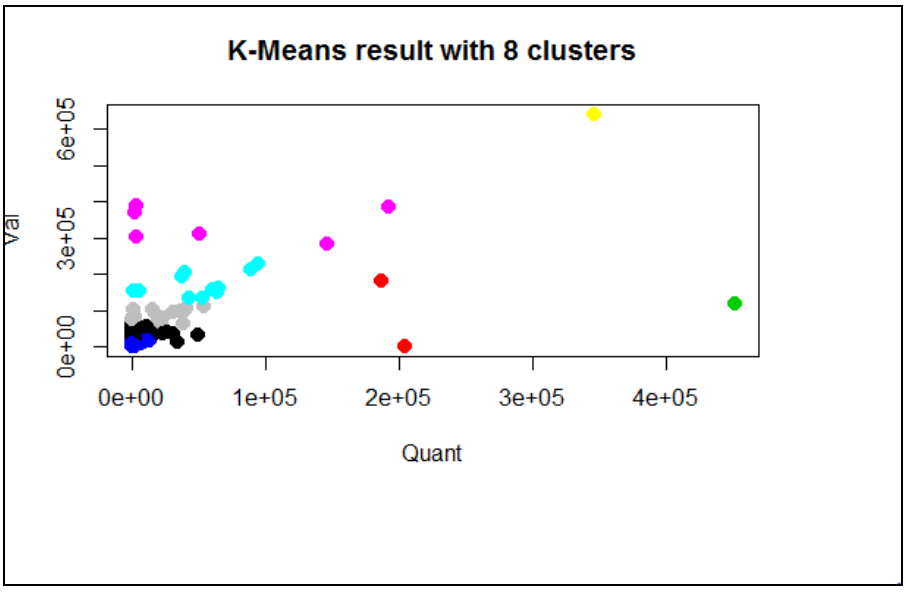


Figure 4. Performance of k-means

Here, for the green dot the transaction involves larger quantity sales, also for yellow dot the quantity as well as value is large which indicates the fraudulent transactions. Pink and red dots represent the unknown values that are likely to be fraudulent. Blue, grey, black and cyan represents the genuine transaction as in Figure 4. The accuracy obtained using this method is 94.5%

On applying the k-modes clustering algorithm, we observe that, in the above diagram, the black dots belong to the group, which has transactions that are not fraudulent. Also, the red dots belong to the group of transactions for which the status of the transactions as to whether the transactions are fraudulent or not and the green dots represent the cluster of the fraudulent transactions as shown in Figure 5. The accuracy obtained by applying the k-modes algorithm is 94.8%

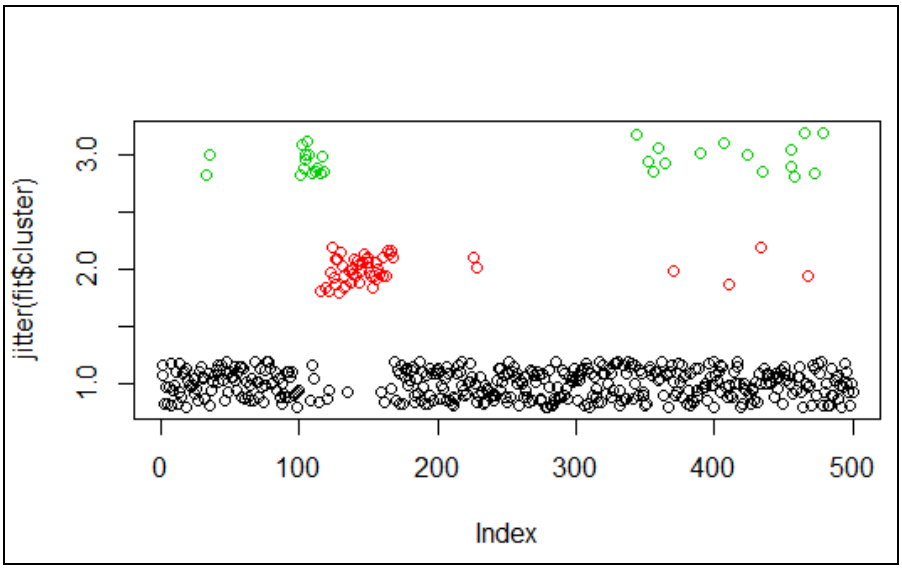


Figure 5. Plot of Data Points in K-Mode Clustering

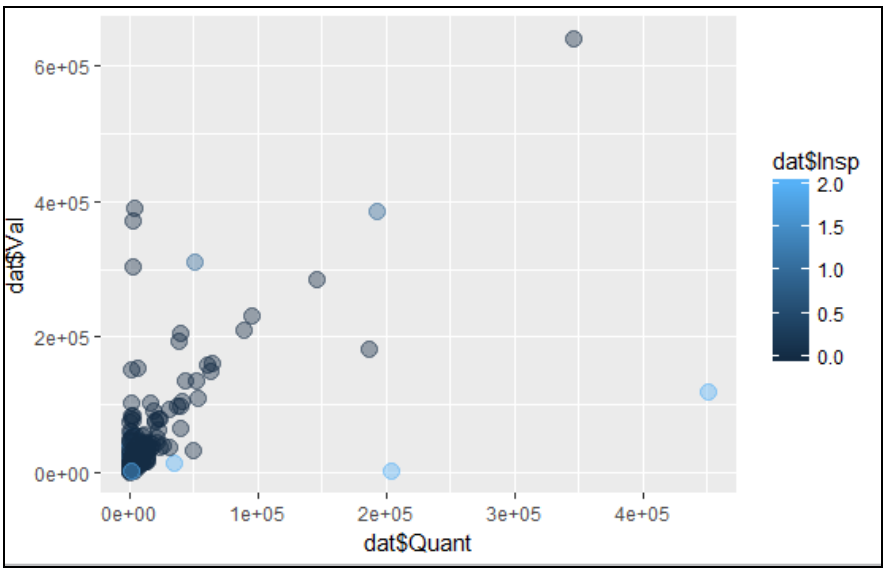


Figure 6. Plot on Hierarchical Clustering

On applying the hierarchical clustering, we observe from the diagram obtained as above, that the dark blue dots represent the transactions that are ok *i.e.*, not fraudulent or not likely to be fraudulent. The medium blue represents the transactions for which the status of the transactions as to whether the transactions are fraudulent or not and the light blue dots represent the transactions that are highly likely, rather, that are fraudulent as in Figure 6. The accuracy obtained by applying the hierarchical clustering is 98.8%

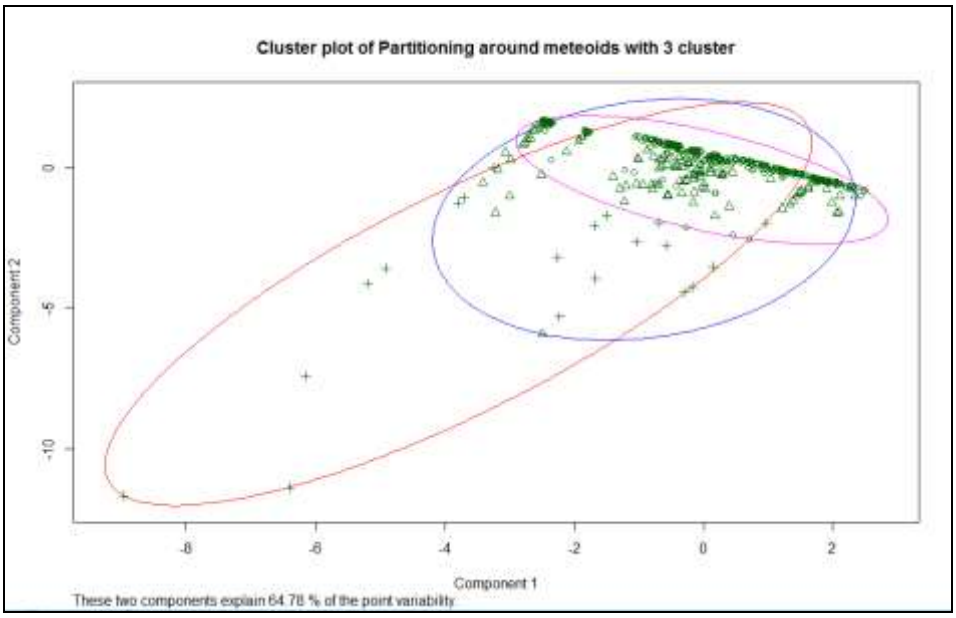


Figure 7. Partitioning around Medoids

On applying partitioning around medoids, we observe that the plus sign indicate the "fraud" values and triangle values indicate the "unkn" values which are the transactions for which the status of the transactions as to whether the transactions are fraudulent or not and circle values are indicating the "ok" values representing the transactions that aren't fraud. The accuracy obtained by applying partitioning around medoids is 94.8 % as in Figure 7.

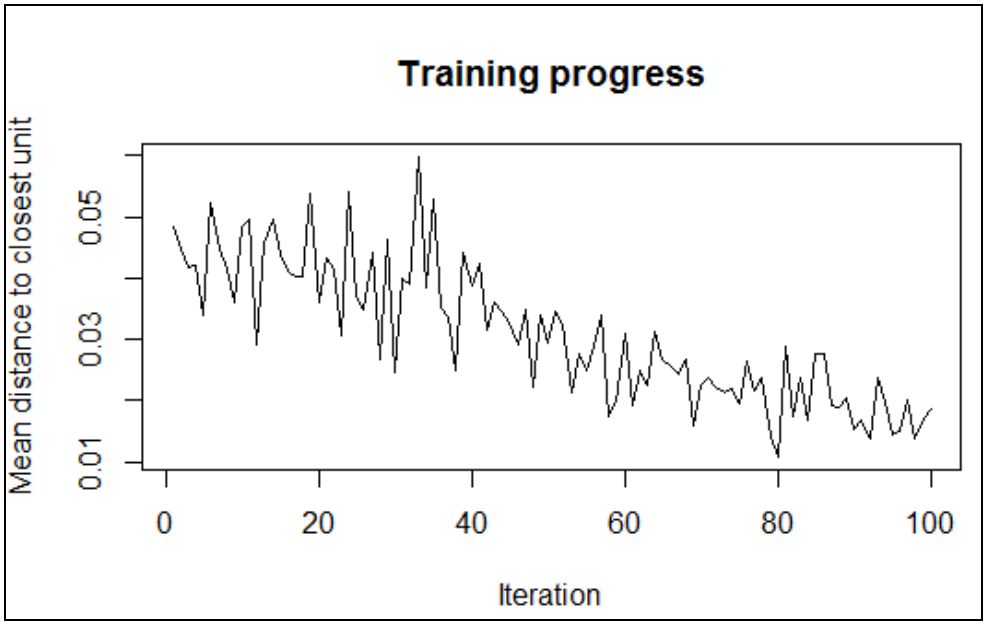
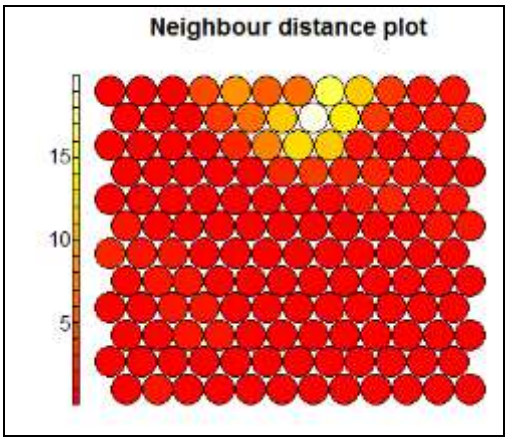
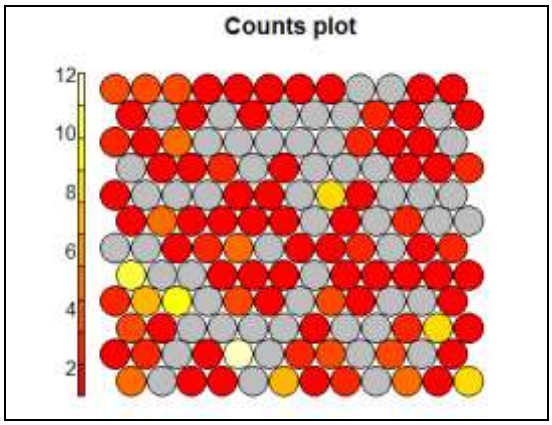


Figure 8. Self-Organizing Map

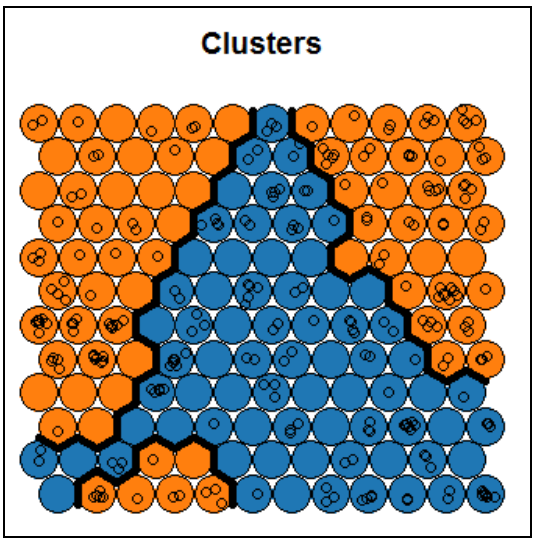
The Self-Organizing map is an unsupervised learning technique that trains on its own.



The Figure 8. obtained as mentioned above shows that as the number of iterations increase the mean distance the closest unit reduces, that is, the similar data points are correctly put together. The diagram above, neighbour distance plot shows that points in red are grouped together and thus the distance between them is less whereas the dots in yellow represent the data points far away.



The counts plotsshow that the red dots are thetransactions for which the status of the transactions as to whether the transactions are fraudulent or not. The orange ones represent the transactions that are not fraudulent. The empty units are denoted by grey units.



Clustering is performed after applying SOM technique. The transactions that are similar to each other are put together. The accuracy obtained by applying the self-organizing map algorithm is 99.2%. Thus, we summarize the observations of the techniques applied as follows:

Table 1. Comparison of Clustering Techniques

Clustering Techniques	Accuracy(in percentage)
K-means algorithm	94.5
K-modes algorithm	94.8
Hierarchical clustering	98.8
Self organising maps	99.2
Partitioning around medoids	94.8

It is observed from the table above that the self-organizing maps give the highest accuracy of 99.2%. Hierarchical also gives good results as 98.8%. K-means algorithm obtains an accuracy as 94.5% and k-modes, partitioning around medoids obtains an accuracy of 94.8%

6. Conclusion

In this paper, we tried and compare different clustering techniques on the sales data for fraud detection. We have applied the k-means, k-modes, hierarchical clustering, Partitioning around medoids, and self-organizing maps. We thus observe that the self organizing maps outperform the other methods by giving an accuracy of 99.2%. Hierarchical clustering also obtains a good accuracy value of 98.8%. Thus, we infer that the self-organizing maps can be efficiently being used for the sales fraud detection.

References

[1] K. Anupriya and C. Kanimozhi, "Predicting Eshopping Data Using Deep Learning", vol. 24, **(2016)**, pp. 250-256.
[2] A. C. Bahnsen, D. Aouada and A. Stojanovic, "Detecting Credit Card Fraud using Periodic Features", **(2015)**, pp. 1-10.
[3] K. Chaudhary and J. Yadav, "A review of Fraud Detection Techniques: Credit Card", vol. 45, no. 1, **(2012)**, pp. 39-44.
[4] Y. Dai, J. Yan, X. Tang, H. Zhao and M. Guo, "Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies", **(2016)**, pp. 251-256.
[5] F. Ghobadi, "Cost Sensitive Modelling of Credit Card Fraud Using Neural Network Strategy", **(2016)**, pp. 8-10.
[6] K. T. Hafiz, S. Aghili and Zavarsky, "The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada", **(2016)**, pp. 1-12.
[7] V. Bhusari and S. Patil, Journal, I., & Applications, C. Study of Hidden Markov Model in Credit Card Fraudulent Detection, vol. 20, no. 5, **(2016)**, pp. 33-36.
[8] P. S. Kalekar and P. Bernard, "Time series Forecasting using Holt-Winters Exponential Smoothing Under the guidance, (4329008), **(2004)**, pp. 1-13.
[9] A. B. Koehler, R. D. Snyder and J. K. Ord, "Forecasting models and prediction intervals for the multiplicative Holt – Winters method", vol. 17, **(2001)**, pp. 269-286.
[10] V. Mareeswari, "Prevention of Credit Card Fraud Detection based on HSVM", (Icices), **(2016)**, pp. 1-4.
[11] H. Modi, S. Lakhani, N. Patel and V. Patel, "Fraud Detection in Credit Card System Using Web Mining", vol. 1, no. 2, **(2013)**, pp. 175-179.
[12] M. P. Namdev, A. Kumar and V. Bansal, "Credit Card Fraud Detection Using an Efficient Enhanced K-Mean Clustering Algorithm", vol. 4, no. 2, **(2015)**, pp. 10367-10374.

[13] A. Salazar, G. Safont, A. Rodriguez and L. Vergara, "Combination of Multiple Detectors for Credit Card Fraud Detection", **(2016)**, pp. 212-217.
[14] J. V. Segura and E. Vercher, "A spreadsheet modeling approach to the Holt ± Winters optimal forecasting, **(2001)**, pp. 131.
[15] M. Vadoodparast and P. A. R. Hamdan, "Fraudulent Electronic Transaction Detection Using Dynamic Kda Model", vol. 13, no. 2, **(2015)**, pp. 32-41.
[16] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", vol. 50, no. 1, **(2003)**, pp. 159-175.
[17] A. Srivastava, "Credit Card Fraud Detection at Merchant Side using Neural Networks", **(2016)**, pp. 667-670.
[18] S. M. Basha, Y. Zhenning, D. Singh Rajput, N. Iyengar and R. Caytiles, "Weighted Fuzzy Rule Based Sentiment Prediction Analysis on Tweets", International Journal of Grid and Distributed Computing, DOI: 10.14257/ijgdc.2017.10.6.04, vol. 10, no. 6, **(2017)**, pp. 41-54.
[19] S. M. Basha, Y. Zhenning, D. Singh Rajput, R. D. Caytiles and N. Ch SN Iyengar, "Comparative Study on Performance Analysis of Time Series Predictive Models", International Journal of Grid and Distributed Computing, DOI: 10.14257/ijgdc.2017.10.8.04, vol. 10, no. 8, **(2017)**, pp. 37-48.
[20] S. M. Basha, H. Balaji, N. Ch SN Iyengar and R. D. Caytiles, "A Soft Computing Approach to Provide Recommendation on PIMA Diabetes", International Journal of Advanced Science and Technology, DOI: 10.14257/ijast.2017.106.03, vol. 106, **(2017)**, pp. 19-32.
[21] S. M. Basha, D. Singh Rajput and V. Vandhan, "Impact of Gradient Ascent and Boosting Algorithm in Classification", International Journal of Intelligent Engineering and Systems (IJIES), DOI: 10.22266/ijies2018.0228.05, vol. 11, no. 1, **(2018)**, pp. 41-49.
[22] S. M. Basha, D. Singh Rajput, N. Iyengar and R. Caytiles, "A Novel Approach to Perform Analysis and Prediction on Breast Cancer Dataset using R", International Journal of Grid and Distributed Computing, http://dx.doi.org/10.14257/ijgdc.2018.11.2.05, vol. 11, no. 2, **(2018)**, pp. 41-54.
[23] V. P. Khadse, S. M. Basha, N. Iyengar and R. Caytiles, "Recommendation Engine for Predicting Best Rated Movies", International Journal of Advanced Science and Technology, http://dx.doi.org/10.14257/ijast.2018.110.07, vol. 110, **(2018)**, pp. 65-76.
[24] S. Dutta, S. M. Basha, N. Iyengar and R. Caytiles, "Classification of Diabetic Retinopathy Images by Using DeepLearning Models", International Journal of Grid and Distributed Computing, http://dx.doi.org/10.14257/ijgdc.2018.11.1.09, vol. 11, no. 1, **(2018)**, pp. 89-106.
[25] D. Khaturia, S. Muzamil, N. Iyengar and R. Caytiles, "A Comparative study on Airline Recommendation System Using Sentimental Analysis on Customer Tweets", International Journal of Advanced Science and Technology, http://dx.doi.org/10.14257/ijast.2018.111.10, vol. 111 **(2018)**, pp. 107-114.
[26] S. M. Basha and D. Singh Rajput, "Evaluating the Impact of Feature Selection on Overall Performance of Sentiment Analysis", In Proceedings of the 2017 International Conference on Information Technology, ACM, **(2017)**, pp. 96-102.
[27] S. M. Basha and D. Singh Rajput, "Fitting a Neural Network Classification Model in MATLAB and R for Tweeter Data set", Proceedings of International Conference on Recent Advancement on Computer and Communication, Springer, Singapore, **(2018)**, pp. 11-18.

