International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

CREDIT CARD DECEPTION DETECTION USING EM ALGORITHM

Jeevan CS[1], Aishwarya B Shetty[1], Shreenidhi NV[1], Apoorva GK[1], Girish SC[2]

[1]UG Student, Dept. of Information Science and Engineering, NIE Institute of Technology, Mysuru, Karnataka, India.
[2]Assistant Professor, Dept. of Information Science and Engineering,NIE Institute of Technology, Mysuru, Karnataka, India.

Abstract - *Credit Card Fraud can be defined as where a person uses other person's credit card for personal reasons, while the owner and the Credit Card are unaware of the fact that the card is being used. Due to the rise of E-Commerce, there has been a high use of credit cards for their personal use which led to High amount of frauds using credit cards. In the era of digitalization, it is necessary to identify credit card fraud. Fraud detection involves observing and analyzing the behavior of various credit card holders to detect or avoid unwanted behavior using credit card. To identify credit card fraud, we need to understand different technologies, algorithms and types involved in detecting credit card frauds. The algorithm can differentiate transactions that are fraudulent or not. Machine learning algorithms are used to analyze all the transactions and report the suspicious ones. In this paper we have done data analysis and based on the analyzed data, it is then passed to different clustering algorithms such as k-means, DBSCAN, spectral clustering, agglomerative clustering and Gaussian Mixture using Exception Maximization algorithm, and the result shows that EM algorithm provides best accuracy.*

Key Words: **Credit card fraud, applications of machine learning, EM algorithm, clustering algorithms, local outlier factor, automated fraud detection.**

1. INTRODUCTION

The Credit Card Fraud Problem includes modeling the past credit card transactions, while knowing the transactions of the ones that turned out to be fraud. The model is then used to identify and detect whether a new transaction is genuine or fraud. Our aim here is to detect the fraudulent transactions by applying different algorithms. We used the implementation of EM Algorithm on Credit Card Fraud data set.

'Fraud' in credit card transactions is unauthorized usage of an account by someone who is not the owner of that account. Necessary measures can be taken to stop this abuse and unwanted behavior of such fraudulent practices. In other words, Credit Card Fraud can be defined as a scenario where a person robs and uses someone else's credit card for their personal reasons, while the owner of the credit card is unaware of the fact that their card is being used by someone.

Fraud detection involves monitoring the activities of different transactions of the users in order to estimate or avoid objectionable behavior, which is fraud, intrusion, and defaulting. By using the solution of Machine Learning such problems can be automated. This problem is particularly very challenging from the perspective of learning, because it is characterized by various factors such as imbalance in the class. The number of genuine transactions is very much greater than that of the known transaction. Also, the transaction patterns change regularly with their statistical properties over the course of time.

Fraud detection based on the analysis of present purchase data of cardholder is a way to reduce the rate of credit card frauds. Since every human being tend to exhibit specific behaviorist profiles, every cardholder can be identified by a different set of patterns containing information about the different purchase category, the time since the last purchase, the amount of money spent on the purchase, etc.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning is to most importantly understand the structure of data and then fit that data into different models that can be understood and utilized by people.

What is E-M algorithm?

Expectation-maximization (**EM**) **algorithm** is an alternative method to find maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables.

What are the evaluation metrics?

The metrics for classification problems mainly come down to:

1. Accuracy
2. Recall and Precision

1. Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Where TP = True Positive, TN = True Negatives, FP = False Positives and FN = False Negatives.

2. Recall and Precision:

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

2. METHODOLOGY

a. Down Sampling:

It is the process of getting the required ratio of labels, since the original data has a skewed dataset. Here the whole of data is sample to suit the ratio of 50:50 labels.

b. Standard Scalar:

It is process of normalization in which the original data is subtracted by mean and divided by the standard deviation. This is done to due to distance metric used in clustering is going to be affected due to different scales of data.

2.1 DATASET

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

2.2 ALGORITHM

Input – Data containing the transactions to be tested against.
Output – Detection of Deception by appropriate labels

Step 1: Accept data from the user containing the transactions to be analyzed.

Step 2: Pre-processing of the data; Standard / Normalization by standard deviation

Step 3: Down-Sample the data for training.

Step 4: Prediction of the label

Let C (n) be set of Extracted Images and
If<0>
Then output -> “Not Fraud”

Else
output - > “Fraud”

Step 5: Stop.

3. PROPOSED SYSTEM

At the start of the method we lay out the data properties and the distributions which gave us an idea of the data and its properties present. Later we drew a graph representing the various correlations of the

variables with each other in the dataset. The proposed system has a technique where since the original data is highly skewed; we down sampled the data in order to balance the ratio of fraud and not fraud labeled data.

Here we are focusing on using E-M algorithm based Gaussian Mixture as our main algorithm for classifying our data along with that we provide a holistic result using other clustering methods like K-means, DBSCAN, algometric etc. The data after being down sampled was subject to normalization after which it was fitted to the models. The hyper parameters of the models where assigned by evaluation of the fit data on the model and measuring the accuracy of the model against the validation data. The best parameter was chosen by this method.

The UI or the web app was developed using FLASK-python framework where in which we used HTML/CSS JavaScript to create the front-end. The application mainly takes in a csv file which contains the test file that we want to check against. After submission the backend would directly Preprocessed the data in the file and run it on the Gaussian-mixture algorithm which was basically an object of the fit data serialized by pickling.

After the data was predicted by the algorithm the output would be sent in the form of a table which has the index of the sample as given in the "csv" file and another column would be the prediction the algorithm had done, in the algorithm standpoint the label was given as binary output and later this binary output was given as "Fraud" or "Not Fraud" based on the given output.

3. RESULTS

i. Algometric Clustering

```
correct positives or frauds predicted 15 out of  57  for test data
```

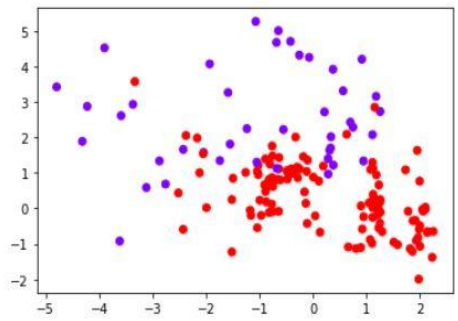

Fig 1: Scatter plot of Algometric Clustering result on the data

Recall Score 26.3% i.e 15 out of 57 were predicted as positive out of the overall set of 57 positive for fraud.

ii. Spectral Clustering

```
correct positives or frauds predicted 1 out of  57  for test data
```

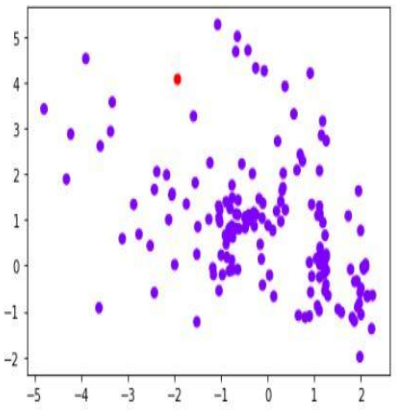

Fig 2: Scatter plot of Spectral Clustering result on the data

Recall Score 1.75 % i.e 1 out of 57 were predicted as positive out of the overall set of 57 positive for fraud.

iii. DBSCAN

```
Estimated number of clusters: 2
Estimated number of noise points: 21
```

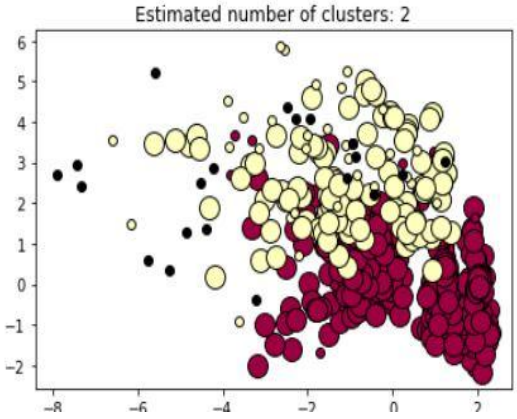

Fig 3: Scatter plot of DBSCAN Clustering result on the data

Recall Score 35.08 % i.e 20 out of 57 were predicted as positive out of the overall set of 57 positive for fraud.

iv. K-Means Clustering

```
correct positives or frauds predicted 40 out of  57  for test data
```

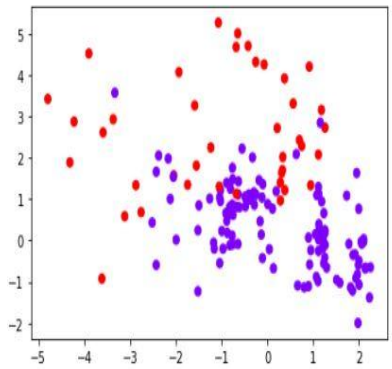

Fig 4: Scatter plot of K-Means Clustering result on the data

Recall Score 70.1 % i.e 40 out of 57 were predicted as positive out of the overall set of 57 positive for fraud.

v. Gaussian Mixture

```
correct positives or frauds predicted 44 out of  57  for train data
```

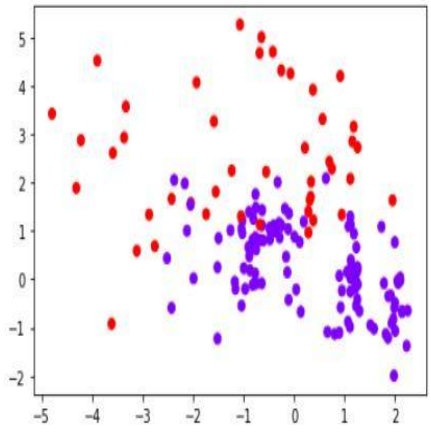

Fig 5: Scatter plot of Gaussian mixture Clustering result on the data

Recall Score 77.1 % i.e 44 out of 57 were predicted as positive out of the overall set of 57 positive for fraud.

The recall score calculated for all the algorithms are as given. Here recall is the main measure used as the percentage of covering the positive labels which count for the fraud label is much more important.

Here the dataset used of testing contains around 159 down sampled points from the overall dataset of 284807 of which 57 are known fraud data.

	Algorithms	Recall Score in percentage
1	Spectral Clustering	1.75
0	Algometric Clustering	26.37
2	DBSCAN	35.08
3	K-Means	70.17
4	Gaussian Mixture	77.19

Table 8.1 Comparison of results

4. CONCLUSIONS

The proposed system suggests that the advanced approach is a worth, which can distinctly support an accurate analysis of the transactions in a minor computational effort. It also dedicates future study on automatically estimating the diagnosis of the transactions. The proposed method for detection of labels has been successful in recognizing and classifying the transactions. The problem with the dataset is that the content of fraud transactions is very skewed hence down sampling is the method to go about, the proposed method can be improved by using bigger data set containing the fraud data. The overall accuracy for classification with the proposed method is 77.1%, which was obtained when minimum distance criterion with Gaussian Mixture clustering had been used. With very less computational efforts, the optimum results were obtained, which also shows the efficiency of proposed algorithm in recognition and classification of the leaf diseases. From the results it can be seen that only few samples were misclassified.

REFERENCES

[1] Nassar, Nader, and Grant Miller. "Method for secure credit card transaction." 2013 International Conference on Collaboration Technologies and Systems (CTS). IEEE, 2013.

[2] Kho, John Richard D., and Larry A. Vea. "Credit card fraud detection based on transaction behavior." TENCON 2017-2017 IEEE Region 10 Conference. IEEE, 2017.

[3] Bahnsen, Alejandro Correa, et al. "Detecting credit card fraud using periodic features." 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2015.

[4] Devi, Debashree, Saroj K. Biswas, and Biswajit Purkayastha. "A cost-sensitive weighted random forest technique for credit card fraud detection." 2019 10th international conference on computing, communication and networking technologies (ICCCNT). IEEE, 2019.