Credit card Fraud Detection

Group 3

Significant Paper Report

Submitted To:
Dr. Vishnu Pendalya

Submitted By:
Mahesh Reddy Konattham
Mohdi Habibi
Rakesh Amireddy
Saumil Shah
Shivangi Nagpal

Paper Studied

John, O. A., Adebayo, O. A., Samuel, A. O. Credit card fraud detection using machine learning techniques: A comparative analysis, ICCNI 2017: International Conference on Computing, Networking and Informatics, Lagos, Nigeria.

Report

Financial fraud is on the rise in the financial industry. With the rise in credit card transactions, credit card fraud rate has also increased. The traditional fraudulent detection is not efficient and takes up a lot of time. So, financial institutions have started implementing computational methodologies. Data mining is a very essential method to solve credit card fraud detection. It basically classifies whether a transaction is legitimate or not. It analyzes the spending behavior, to detect the problem. In this paper, data mining techniques are implemented on a dataset. Different algorithms are then compared and evaluated using several metrics to decide which technique works the best. But there are a lot of issues in implementing this, such as, Fraudulent behavior can vary and can look like a legitimate transaction. There are not many datasets available and they are altered and imbalanced making it difficult to work on. This paper analyzes the performance of naïve bayes, k-nearest neighbor and logistic regression on a highly imbalanced dataset containing information about European card holders to detect fraudulent acts. A comparative study is carried out based on metrics like accuracy, sensitivity, specificity and Matthew's correlation coefficient. Since the data is imbalanced, the labels are oversampled and undersampled accordingly, upon which the algorithms are applied.

There are different categories of fraudulent acts and with time, fraudsters learn to mimic legitimate user behavior, making it difficult to differentiate the normal and fraud behavior. Here, three different sample ratios are used to get the best sample performance. Stratified sampling is applied on the dataset to not disturb and alter the patterns in the data. Data mining strategies analyze the user behavior to detect a fraud. The optimal selection of features is a must to classify legitimate and a fraudulent transaction. In this study, they have focused on card details compared to card holder details, to capture the behavior. The variables are classified as all transactions statistics, regional statistics, merchant type statistics, time- based amount statistics and time-based number of transactions statistics.

For the experiment, the dataset was of transactions made by European people in September 2013 that took place in 2 days, which includes 284807 transactions. The dataset is highly skewed towards positive class. After applying Principal Component Analysis, 28 principal components were deduced. The dataset is both under sampled and over sampled, thus giving 2 distributions of dataset, upon which the algorithms are applied.

The first algorithm implemented was Naïve Bayes, which is based on Bayesian theory, making a decision based on highest probability. Naïve Bayes makes a conditional independent assumption for the variables and then performs binary classifications based on Bayesian classification rule. The next algorithm implemented was k-nearest neighbor. It is an instance-based learning which classifies based on similarity measures and can work on continuous and categorical data using different distance measures. For this study, Euclidean distance is considered, which is the distance between and input data point and current point. Using this distance, k-nearest neighbors are found, and those points are grouped in a cluster. In this study, they have chosen k = 3, as it gave the most optimal performance. Lastly, they implemented Logistic Regression Classifier, which helps to find the best fit parameters to sigmoid function. It rounds the data between 0 and 1. To evaluate the performance of the classifier, the gradient ascent and modified stochastic gradient ascent optimization methods were used. After 100 iterations, stable values of parameters were achieved. In this experiment, due to the size of the data, gradient ascent is used, to reduce computational complexity.

After implementing the algorithms, the important metrics for evaluation are True positive, True negative, False positive, False negative, which will help to find the evaluation measures to find the performance of individual algorithms. In this study, Sensitivity gave accuracy on positive cases classification; Specificity gave accuracy on negative cases classification; Precision gave accuracy on positive cases classification; MCC was used on an unbalanced data. The data was divided into 70% training and 30% test sets. According to the evaluation, k-NN outperformed Naïve Bayes and Logistic Regression Classifier for all the evaluation metrics, i.e., Accuracy, Specificity, Sensitivity, MCC and balanced classification rate. The lowest performance of k-NN recorded was for the 10:90 data distribution.

To conclude, this study depicts the effect of hybrid sampling on the performance of binary classification of skewed data. The expected future work is analyzing eta classifiers and meta learning strategies in approaching highly skewed credit card dataset and effects of other sampling methods.