Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol.6, No.2, December 2018, pp.394-400
Available online at:http://pen.ius.edu.ba

Using Isolation Forest in anomaly detection: the case of credit card transactions

Soumaya Ounacer[1], Hicham Ait El Bour[1], Younes Oubrahim[1], Mohamed Yassine Ghoumari[1] and Mohamed Azzouazi[1]

[1] Department of mathematics and computer science. Information Technology and Modeling Laboratory. Hassan II University, Faculty of Sciences Ben M'Sik Casablanca, Morocco.

Article Info

Received Jul 5, 2018

Keyword:

Anomaly detection
Credit card frauds
LOF
OCSVM
K-means
Isolation Forest

ABSTRACT

With the evolution of new technology especially in the domain of e-commerce and online banking, the payment by credit card has seen a significant increase. The credit card has become the most used tool for online shopping. This high rate in use brings about fraud and a considerable damage. It is very important to stop fraudulent transactions because they cause huge financial losses over time. The detection of fraudulent transactions is an important application in anomaly detection. There are different approaches to detecting anomalies namely SVM, logistic regression, decision tree and so on. However, they remain limited since they are supervised algorithms that require to be trained by labels in order to know whether the transactions are fraudulent or not. The goal of this paper is to have a credit card fraud detection system which is able to detect the highest number of new transactions in real time with high accuracy. We will also compare, in this paper, different unsupervised techniques for credit card fraud detection namely LOF, one class SVM, K-means and Isolation Forest so as to single out the best approach.

Corresponding Author:

Soumaya Ounacer,
Department of mathematics and computer science, Information Technology and Modeling Laboratory,
Hassan II University, Faculty of Sciences Ben M'Sik Casablanca, Morocco.
Email: soumayaounacer@gmail.com

1. Introduction

Given the high number of purchases made daily using the credit card, the risk of fraudulent transactions arises at any time. Fraudsters are in strong improvements to their fraud models; they change their behavior all the time. In order to recognize fraudulent transactions made by the credit card, we will apply anomaly detection. Anomaly detection is an important problem in a wide range of application domains and research areas such as health diagnosis, system intrusion in cyber-security, credit card fraud detection, e-commerce[1]and fault tolerance in industry and so on.

An anomaly refers to when something substantially deviates from the normal behavior, and detecting such anomalies in data is called anomaly detection[2].Anomaly detection is a technique which determines all such instances that deviates from an expected behavior[3]. Many anomaly detection methods, supervised and unsupervised, have been applied to credit card fraud. The supervised techniques such as SVM, Decision trees, KNN, logistic regression and others give better results and can solve the problem of detecting fraud to some extent. [4]Yet, these methods require labeled data to form the classifier with fraudulent and non-fraudulent

behaviors whereas in the unsupervised techniques the data does not have to be labeled. It is based on the premise that fraudulent behavior will act very differently than normal.
The majority of works in detecting fraudulent transactions of credit card deals with supervised techniques[5]. Finding enough examples of fraudulent transactions is a big challenge. In real situations, it's very difficult to obtain labeled datasets especially in the area of anomaly detection and when the labeling is performed by humans this operation is expensive. The efficiency of supervised algorithms may be affected because of dealing with a heavily imbalanced class distribution, which leads to anomaly detection [6]. For that we will focus on the unsupervised anomaly detection approach.
The objective of this work is to identify whether new transactions are fraudulent or not by using the isolation forest technique which helps in minimizing the number of false positives and detecting the highest number of fraud in credit card transactions. This paper is organized as follows: In the first section, we will discuss some related works in the case of credit card fraud and then, in the next section, we will compare different unsupervised methods that identify anomalies in the credit card transactions. The adapted model will be explained thoroughly in section three. In the fourth one, we shall demonstrate how the experiment results validate our proposed technique. Last but not least, we will conclude this paper with a conclusion and give some ideas for future work.

2. Related Works

There are several works in the field of credit card anomaly detection based on the DataMining approach[7]. Maes et al. [8]have discussed the problem of credit card fraud detection using Bayesian Belief Network and Artificial Neural Network. The comparison of the two machine learning techniques shows that Artificial Neural Network detects the fraudulent transactions much faster than the Bayesian Belief Network. This latter, displays better results; it detects 8% more of the fraudulent transactions than the other.
A comparative study was conducted on both decision tree and Support Vector Machine techniques on a data set containing credit card transactions[9]. Decision tree method caught up to 33% more frauds than Support Vector Machine. The two methods have similar detection accuracy.
The K-nearest neighbor is a supervised learning technique used to detect credit card frauds. It calculates its nearest point in order to classify credit card fraud detection. [10] K-nearest neighbor recognizes this transaction as a fraud if the new transaction is coming and the point is near the fraudulent transaction. This technique cannot detect the anomaly at the time of transaction.

3. Credit card fraud detection techniques

Anomaly detection in the case of credit card fraud is divided into supervised and unsupervised techniques.
The supervised methods have many drawbacks. If a fraudulent transaction happen which is not conformed to the database, these transactions will be considered normal whereas with unsupervised methods, anomaly patterns are found through new transactions and occurred information.
In this section we compare different unsupervised techniques in order to choose the most adequate one to detect real time anomalies in credit card transactions.

- Local Outlier Factor (LOF)[11] Was proposed by breunig et al, is an unsupervised anomaly detection algorithm. In each point, it computes the density of its local neighborhood. One can recognize regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers. LOF shares some concepts with OPTICS and DBSCAN which are used for local density estimation such as the concepts of "reachability distance" and "core distance"[12].
- One-Class Support Vector (OCSVM) is an unsupervised anomaly detection algorithm. [13] It was proposed by Scholkopf to identify anomaly without labeled training data. It is an application of support vector machine algorithms to one class problems. This method separates by a hyperplane all the data points from the origin. It estimates the density level sets and gives an estimate of underlying density.
- K-means (KM) [14]is an unsupervised algorithm for anomaly detection. It divides data into K-clusters and guarantees that the data in various clusters have low similarity, while the data within the same cluster are similar. This clustering algorithm evaluates similarity and calculates the distance between two samples. It is an efficient method to cluster the data.

Table 1 represents a comparison of three anomaly detection techniques' advantages and disadvantages.

Table 1. Comparison of Unsupervised Anomaly Detection Techniques

Anomaly Detection Techniques	Advantages	Drawbacks
One class SVM	Finds the good separation hyperplane Works well	Not suitable for large amounts of datasets Needs lots of memory Some numerical stability problems
Local Outlier Factor	Good algorithm for local anomaly detection	Only relies on its direct neighborhood Performs poorly on datasets with global anomalies Scalability is a big issue
K-means	Easy to implement Low complexity	Sensitive to noise and outlier data points Each cluster has pretty equal numbers of observations Clusters are sensitive to initial assignment of centroids Necessity of specifying K Only work with numerical data

4. Isolation Forest

Under the umbrella of anomaly detection lays one of the most studied branches; credit card fraud detection. Most of the well known anomaly detection approaches create, first, a profile of normal instances and then identify whatever does not fit into the normal profile as an anomaly [15]. The proposed Isolation Forest in Liu et al. (2008) 'isolates' observations selecting an attribute and then selecting a split value between the maximum and minimum values of the selected attribute in an arbitrary way. The number of splittings required in order to isolate a sample is equivalent to the path length from the root node to the terminating node[16]. because the recursive partitioning can be represented by a tree structure. This path length which is averaged over a forest of random trees is a scale of abnormality. The averaged depth is the basis of the scoring function. As shown in figure 1, random partitioning creates shorter paths for anomalies. The average depth of a sample over the forest seems to meet, to some extent, the latter being different if the sample is or is not an anomaly.

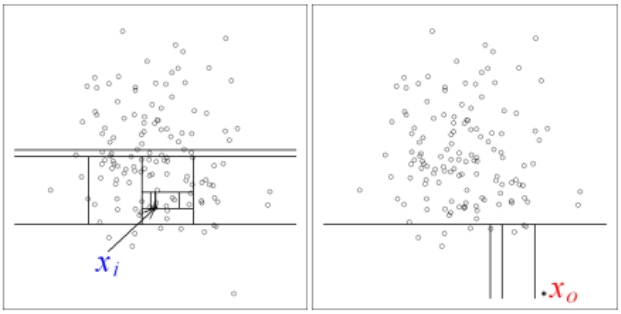

Figure 1. Identifying Normal vs. Abnormal

Many advantages make Isolation forest outrank other methods in anomaly detection algorithm. First, it only needs small samples from large datasets so as to derive an anomaly detection function which makes it fast and scalable. Second, it does not require example anomalies in the training dataset. Third, the tree depth is the basis of its distance threshold for determining anomalies which is autonomous from the scaling of the dataset dimensions. It can both work as a supervised and an unsupervised classifier and its objective is that the anomalies are less recurrent than usual observations and vary from their values. For that, applying such arbitrary partitioning, requires a closer identification to the tree root with less conditions needed. However, isolating normal data points needs more conditions. Figure 1 clearly shows this concept of identifying a

normal versus an abnormal observation. It shows that a normal point x_i requires more divisions in order for it to be isolated. Conversely, an anomaly x_0 needs fewer divisions to be isolated. Consequently, an anomaly score can be calculated as the number of conditions required to isolate a given observation. There are only two training parameters, which are the number of trees to build and sub-sampling size, in addition to one evaluation parameter which is the tree height limit during evaluation in this method[17].
Isolation Forest calculates an anomaly score for decision making. It is defined as:

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}}$$

E(h(x)) is the average of h(x) from a collection of isolation trees. In this equation, when $E(h(x)) \rightarrow c(\psi)$, $s \rightarrow 0.5$ that is when the S of data return is very close to 0.5, then in the entire sample there is no distinct anomaly. When $E(h(x)) \rightarrow 0$, $s \rightarrow 1$ that is when the S is very close to 1, they indicates anomalies. And when $E(h(x)) \rightarrow \psi - 1$, $s \rightarrow 0$ that is when the S of the data is much smaller than 0.5 then they indicate a big potential to be rated as normal instances.

5. Proposed Model

In this section we will build an anomaly detection framework that is able to pre-processing, training, predicting data transactions in real time.
The adapted model using isolation forest as follows: When new transactions arrive, the iForest algorithm is occupied in order to assign a score to the observation that will clarify whether the transaction is fraudulent or not. Anomaly score close to 0 is considered normal, value 1 is considered as an anomaly (figure 2).

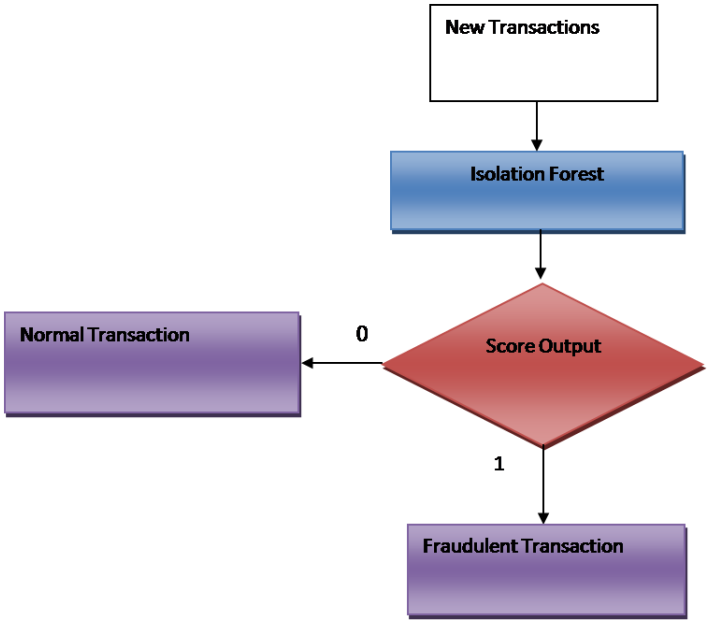

Figure 2.Workflow of Credit Card fraud Detection

Anomaly detection using Isolation Forest has two stages:
- Training stage: builds IForest, we start to multiple isolation trees using subsamples of the training set.
- Testing Stage: passes each data point through isolation tree to calculate the average number of splits across all the trees that isolate observation in order to obtain an anomaly score for each instance[18].
The figure 3 shows the overall flow of credit card anomaly detection. The first step is to build the training model from incoming credit card data transactions. Whereas the second step, when new transactions arrive, it predicts data transactions using its model training.

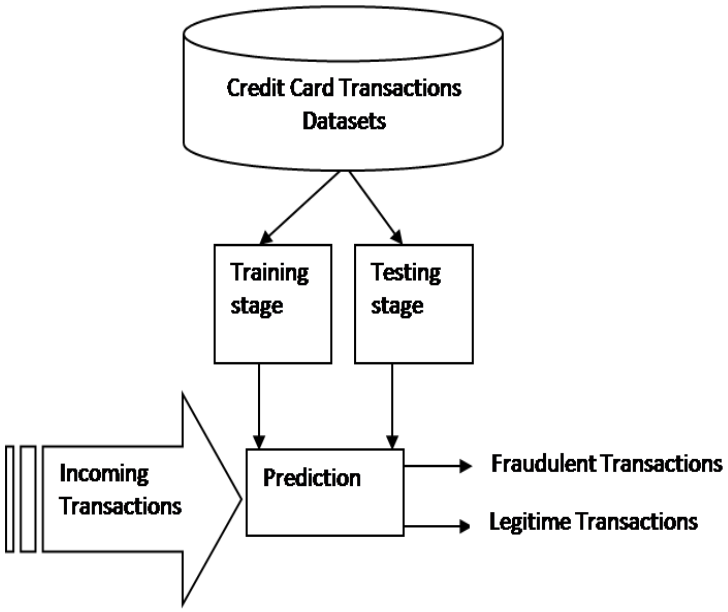

Figure 3. Credit Card fraud detection Model

6. Experiments

The credit card fraud datasets, which were taken from Kaggle[19], are from a European credit card company. It contains 2 days transactions which were previously made by the credit card holder in September 2013. The analyzed data contains just 492 frauds out of 284,807 transactions (0.172%).

Due to confidentiality matters, all input variables are converted to numerical value by PCA transformation. The original features are not presented and features V1, V2, ... V28 are the principal components obtained after the transformation and the features that are not PCA transformed are 'Time' and 'Amount'. The first represents the difference, in seconds, between the particular transaction and first transaction. 'Amount', on the other hand, represents the money transaction that had occurred, feature 'Class' is a variable that shows if a transaction is fraudulent or not. If value is 1, it shows that the transaction is fraud or else it is not fraudulent.
This experiment has helped us choose the most appropriate approach in the case of detecting credit card anomalies. A comparison was made taking into consideration several performance measures namely, the rate of false positive, accuracy, error rate ...etc. The results of the four models are shown in the following table and confusion matrice.

Table 2.Performance Mesure

Methods	F1 Score	Accuracy	AUC Score
OCSVM	0,0033	0,5088	0,5154
LOF	0,0027	0,8901	0,4970
K-means	0,0054	0,9012	0,5191
Isolation Forest	0.0544	0.9512	0.9168

Figure 4. Confusion Matrix using IForest

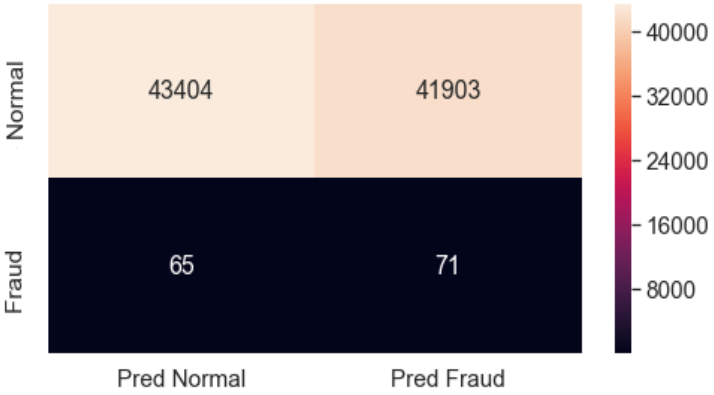

Figure 5. Confusion Matrix using OCSVM

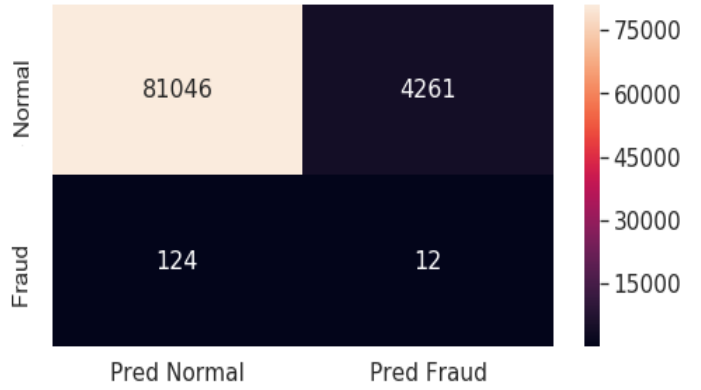

Figure 6. Confusion Matrix using K-means

Figure 7. Confusion Matrix using LOF

Isolation Forest algorithm performs well in the case of credit card. Isolation forest has a 91% AUC score than LOF 49%, OCSVM 51% and K-Means 52%. This algorithm is more accurate than others and it detects few errors than the three other methods. Isolation Forest performed much better in detecting fraudulent transactions.

7. Conclusion

In this paper, we have proposed to use isolation forest for detecting fraudulent transactions of credit card. This choice was made through evaluating different methods including OCSVM, LOF and K-means in terms of accuracy, F1score, AUC score and false positive rate of the system.

Experimentation has shown that the isolated forest is very efficient in detecting anomalies in the case of the credit card. In future work, we will implement a new architecture capable of detecting fraudulent transactions in real time by combining Apache spark and isolation forest.

References

[1] T. Banerjee, M. Mishra, N. C. Debnath, and P. Choudhury, "Implementing E-Commerce model for Agricultural Produce : A Research Roadmap," vol. 7, no. 1, pp. 302–310, 2019.

[2] HuaMing, Kishan G. Mehrotra Chilukuri K. Mohan, Anomaly Algorithms Principles and Detection.2017.

[3] V. CHANDOLA, A. BANERJEE, and V. KUMAR "Anomaly Detection: a SURVEY," Conform. Predict. Reliab. Mach. Learn. Theory, Adapt. Appl., vol. 41, no. 3, pp. 71–97, 2014.

[4] R. A. Flarence, S. Bethu, V. Sowmya, K. Anusha, and B. S. Babu, "Importance of Supervised Learning in Prediction Analysis," vol. 6, no. 1, pp. 201–214, 2018.

[5] N. Venkateswaran, A. Shekhar, and S. Changder, "Using machine learning for intelligent shard sizing on the cloud," vol. 7, no. 1, pp. 109–124, 2019.

[6] D. Sabinasz, "Dealing with Unbalanced Classes in Machine Learning," Deep Ideas. 2017.

[7] H. Jihal, M. A. Talhaoui, A. Daif, and M. Azzouazi, "Predictive Analytics as A Service on Moroccan Tax Evasion," vol. 7, pp. 90–92, 2018.

[8] S. Maes and K. Tuyl, "Credit Card Fraud Detection Using Bayesian and Neural Networks," no. August 2002, 2013.

[9] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines," vol. I, 2011.

[10] N. Malini and M. Phil, "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection," pp. 3–6, 2017.

[11] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF : Identifying Density-Based Local Outliers," pp. 1–12, 2000.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "OPTICS-OF: Identifying Local Outliers," no. September, pp. 262–270, 2010.

[13] B. Sch and R. Williamson, "Support Vector Method for Novelty Detection," pp. 582–588, 2000.

[14] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," Expert Syst. Appl., vol. 40, no. 1, pp. 200–210, 2013.

[15] E. Lewinson, "Outlier Detection with Isolation Forest," Towards Data Science. 2017.

[16] F. T. Liu and K. M. Ting, "Isolation Forest."2009.

[17] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 413–422, 2008.

[18] F. T. Liu and K. M. Ting, "Isolation Forest: Isolation Forest," 1980.

[19] "Credit Card Fraud _ Kaggle, Anonymized credit card transactions labeled as fraudulent or genuine." .