IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VI (Mar-Apr. 2014), PP 44-48
www.iosrjournals.org

A Survey on Outlier Detection Techniques for Credit Card Fraud Detection

Ms. Amruta D. Pawar[1], Prof. Prakash N. Kalavadekar[2],Ms. Swapnali N. Tambe[3]
[1]*(Computer Department, SRESCOE Kopargaon, India)*
[2]*(Computer Department, SRESCOE Kopargaon, India)*
[3]*(Computer Department, SRESCOE Kopargaon, India)*

Abstract: *Credit card fraud detection is an important application of outlier detection. Due to drastic increase in digital frauds, there is a loss of billions dollars and therefore various techniques are evolved for fraud detection and applied to diverse business fields. The traditional fraud detection schemes use data analysis methods that require knowledge about different domains such as financial, economics, law and business practices. The current fraud detection techniques may be offline or online, and may use neural networks, clustering, genetic algorithm, decision tree etc. There are various outlier detection techniques are available such as statistical based, density based, clustering based and so on. This paper projected to find credit card fraud by using appropriate outlier detection technique, which is suitable for online applications where large scale data is involved. The method should also work efficiently for applications where memory and computation limitations are present. Here we have discussed one such unsupervised method Principal Component Analysis(PCA) to detect an outlier.*

Keywords : *Clustering, Fraud detection, Neural Networks, Outlier detection, Principal Component Analysis.*

I. INTRODUCTION

1.1 Introduction

Fraud is defined as the use of one’s asset for personal enrichment through misbehavior. In real world fraudulent activity may occur in many areas such as online banking, E-Commerce, mobile communications, telecommunication networks. Fraud is increasing drastically with globalization and modern technology which results in major loss to the businesses. Fraud detection refers to the act of identifying frauds as early as possible. In recent years fraud detection has been implemented using techniques such as neural networks, data mining, statistics, clustering etc. [1].

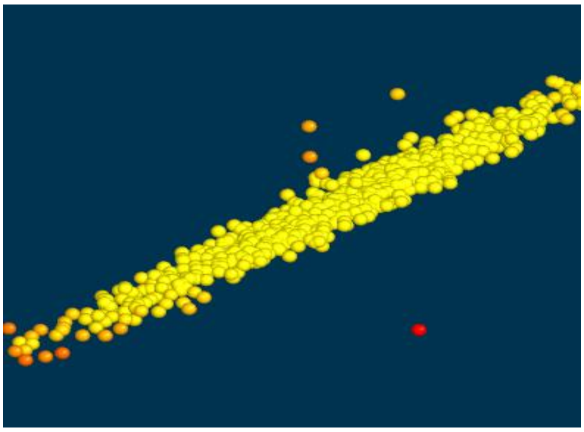

Fig. 1. Concept of an Outlier

An outlier is defined as an observation which deviates so much from other observation. It is also referred to as abnormalities, anomalies or discordance. Fig. 1 shows the concept of an outlier, here all data instances with yellow color are normal data points on the other hand data instance with red color, which is far away from all other data points is an outlier. Outlier detection is a mechanism of detecting an outlier from a given dataset. Outlier detection is applied to various application areas such as homeland security, fraud detection, intrusion detection etc. [2]. In past many outlier detection techniques have been proposed. Statistical based approaches follow some predefined distribution or stochastic model. The data instances that satisfies the distribution model are normal data and the remaining are considered as an outliers. Classification based approaches works in two phase fashion: training phase and testing phase. The training phase learns a classifier and testing phase classifies data instances. Distance based approaches considers distance from neighboring instances and those which are far away are outliers [3].

Outlier detection is an unsupervised data learning problem. The addition or removal of abnormal data instance will affect the principal direction than addition or removal of normal one. This strategy is used to determine the outlierness of the target data instance. In real world application such as credit card fraud detection large amount of data need to be consider and hence here we are exploring different methods of outlier detection [4].

II. RELATED WORK

Y Kou, C. Lu, S. Sinvongwattana and Y. Huang have presented a survey on various fraud detection techniques such as credit card fraud detection, telecommunication fraud detection and intrusion detection system. For each of this category different detection techniques are discussed by them [1].

M. Breunig, H-P Kriegel, R. T. Ng and J. Sander proposed density-based Local Outlier Factor (LOF) method to find outliers. They assigned each object a degree of being an outlier. This degree is called as Local Outlier Factor (LOF) of an object. The degree depends on how object is isolated from surrounding objects. All previous methods consider outliers as binary property but this method assigns degree of being an outlier. However this method does not work on high dimensional dataset and also requires high computation [5].

V. Chandola, A. Banerjee and V. Kumar presented a survey on outlier detection techniques. For each technique they have given normal and anomalous behavior along with its advantages and disadvantages [6].

L. Smith presented all mathematics related to Principal Component Analysis (PCA) from variance, covariance, eigenvalue, and eigenvectors. He also had shown how to calculate PCA by using 2 dimensional data, with scatterplots [7].

Wei Wang, Xiaohong Guan, and Xiangliang Zhang proposed Principal Component Analysis (PCA). This method uses system call data and provides low overhead and high efficiency. PCA is applied to reduce high dimensional data vectors. This method is feasible for real time intrusion detection with high detection accuracy and low computation expenses. This method takes into account frequency property, instead of considering the transition information of the system calls or commands. However research is in progress to mix the frequencies property with the transition information of system calls so that lower false alarms can be achieved [8].

X Song, M. Wu,C. Jermaine, Sanjay Ranka presented a way to determine whether reported anomaly is interesting or not.They make use of domain knowledge provided by user and set of environmental attributes. This conditional anomaly detection technique uses the differences between these attributes and proposes three different expectation-maximization algorithms [9].

C. Aggarwal and P. Yu presented a method to find outliers from high dimensional data space. They used the behavior of projections in order to find out outliers [10].

III. DATASET

To examine outlier detection for credit card fraud we use standard German Credit Card Fraud dataset which is available on UCI Machine Learning Repository. This dataset consist of 20 attributes and 1000 instances. The attributes are both numerical and categorical. The descriptions of attributes are shown in Table1. We also use two dimensional synthetic data to test this technique. In synthetic data set we will generate 100 data instances in which 80 are normal instances and remaining 20 are outliers.

Table 1 Attributes Description

Sr. No.	Attribute	Type	Sr. No.	Attribute	Type
01	Status of existing checking account	Categorical	11	Present residence since	Numerical
02	Duration in month	Numerical	12	Property	Categorical
03	Credit history	Categorical	13	cc_age in months	Numerical
04	Purpose	Categorical	14	Other installment plans	Categorical
05	Credit amount	Numerical	15	Housing	Categorical
06	Savings account/bonds	Categorical	16	Number of existing credits at this bank	Numerical
07	Present employment since	Categorical	17	Job	Categorical
08	Installment rate	Numerical	18	Number of people being liable to provide maintenance	Numerical
09	Personal status and sex	Categorical	19	Telephone	Categorical
10	Other debtors / guarantors	Categorical	20	foreign worker	Categorical

IV. CREDIT CARD FRAUD DETECTION

Outlier detection techniques are applied to detect fraudulent credit card usage. Detecting fraudulent credit card usage is similar to detect an outlier. The data consist of several dimensions such as user id, amount spent, and time between consecutive card usages and so on. The frauds are reflected in transactional records and correspond to high payment, purchase of items that was never purchased by customer, high rate of purchase and so on. The credit card companies have complete data associated with them and have label associated with them. Credit card fraud detection is confidential and it is not disclosed in public. Some methods are discussed here that can be applied to detect credit card frauds [6].

4.1 Outlier Detection

Outlier detection is divided into three main classes based on extent to which labels can be assigned to each group.

4.1.1 Supervised Outlier Detection

In supervised outlier detection mode training dataset is available for both normal as well as outlier classes. This approach builds a predictive model for both these classes and any new data instance is compared against these models. There are certain challenges in supervised outlier detection such as outlier data instances are few as compared to normal data instances. Also it is difficult to obtain the accurate class labels.

4.1.2 Semi-supervised Outlier Detection

In semi-supervised outlier detection mode training dataset is available only for normal class; hence it is widely used than supervised mode. The new target instance is compared against this normal class, the data instances which do not satisfies this class will be consider as an outlier. It is not commonly used because; it is difficult to cover each abnormal behavior to generate normal class.

4.1.3 Unsupervised Outlier Detection

In unsupervised outlier detection mode training data is not available, this technique make assumption that normal data instances are frequent than outliers. The data instances which are frequent or closely related are considered as normal and remaining are outliers. These techniques are widely used as it does not require training data set.

4.1 Neural Networks Based

Neural network based outlier detection technique works in two phases. In first phase neural network is trained to build the normal classes. Second each target instance is tested against those classes by providing input to neural network. If the neural network accepts the data instance it is normal and if the neural network rejects data instance it is deemed as an outlier. Different types of neural networks are derived from basic neural network. Replicator neural network has invented for one class outlier detection. A multilayer feed forward neural network is having same number of input and output neurons. The training phase compresses the data and testing phase reconstructs it.

4.2 Rule Based

Rule based outlier detection technique learns the rule for normal behavior of the system. A test which is not covered by such rule is considered as an outlier. Rule based techniques uses two steps. In first step it learns rules from training data using rule learning algorithm. Each rule has associated confidence value that is proportional to the ratio between number training instances correctly classified by the rule and total number of instances covered by the rule. The second step is to find the rule that best captures the test instance. The inverse of the confidence associated with the best rule is the outlier score of test instance.

4.4 Clustering Based

Clustering is unsupervised outlier detection technique for grouping similar type of data into clusters. Clustering based technique is divided into three categories as follows:

1. Normal data instances belong to a cluster in the data while outlier does not belong to any cluster.
2. Normal data instances are closet to its nearest centroid, while outliers are far away from centroid.
3. Normal data instances belong to large and dense clusters, while outliers belong to small or sparse clusters

V. PRINCIPAL COMPONENT ANALYSIS

5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised dimension reduction method. It is a mathematical algorithm that transforms set of correlated variables into smaller number of uncorrelated variables.

This dimension reduction is performed by identifying directions called principal components which consist of maximum information. By using only few components it is possible to represent each sample with few number of values instead of thousands values. Due to this it is possible to plot samples, visually assess the similarities in order to group data. In order to calculate PCA following steps should be carried out on data matrix.

Step 1: Organize data into pxn matrix, where p is the number of dimensions and n is the number of data instances.
Step 2: Subtract man from each dimension.
Step 3: Calculate covariance matrix.
Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix.
Step 5: Rearrange the eigenvectors and eigenvalues in descending order.
Step 6: Calculate cumulative energy content and select eigenvectors with highest eigenvalues by discarding lower eigenvectors [3].

Let,

$$A = [x_1^T; x_2^T; \dots; x_n^T] \in IR^{n\times p} \qquad \dots (1)$$

Where each row xi represents a data instance in a p dimensional space, and n is the number of the instances. PCA can be calculated by following formula,

$$\underset{U \in IR^{n\times p}}{\text{Max}} \; \|U\| = I\,(U) = \sum_{i=1}^{n} \|(x_i - \mu) - UU^T(x_i - \mu)\|^2 \qquad \dots (2)$$

Where,

$U^T(x_i - \mu)$ Determine the optimal coefficients to weight each principal direction. The PCA problem can be solved by deriving an eigenvalue decomposition problem of the covariance data matrix,

$$\Sigma_A \mathrm{U} = \mathrm{U}^\wedge \qquad \dots (3)$$

Where,

$$\Sigma_A = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \qquad \dots (4)$$

Eq. (4) is the covariance matrix, μ is the global mean. Each column of U represents an eigenvector of ∑A and the corresponding diagonal entry in ^ is the associated eigenvalue. For dimension reduction, the last few eigenvectors will be discarded due to their negligible contribution to the data distribution.

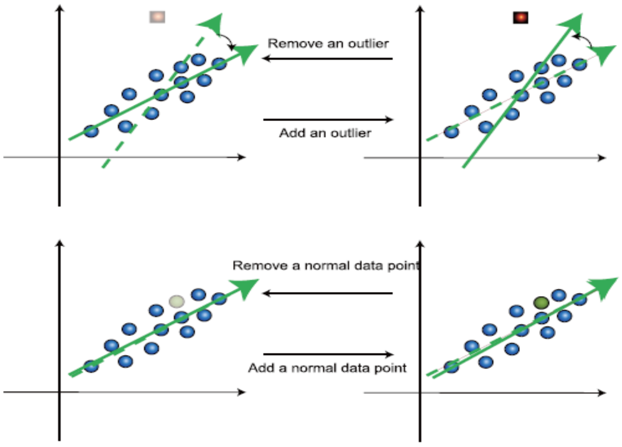

Fig. 2. Effects of adding/removing an outlier or a normal data instance on the principal directions

The blue circles in Fig. 2 represent normal data instances, the red square denotes an outlier, and the green arrow is the dominant principal direction. From Fig. 2, we see that the principal direction is deviated when an outlier instance is added. The presence of an outlier instance produces a large angle between the resulting and the original principal directions. But this angle will be small when a normal data point is added. Therefore, this technique uses this property to determine the oulierness of the target data point [4].

5.2 Use of PCA for Outlier Detection

Given a data set A with n data instances, we first extract the dominant principal direction u from it. If the target instance is a_t we next compute the leading principal direction $\tilde{u}_t$ without a_t present. To identify the outliers in a data set, we simply repeat this procedure n time.

$$\Sigma_{\tilde{A}}\tilde{u}_t = \lambda\tilde{u}_t \qquad \dots (10)$$

Where,

$$\tilde{A} = A \backslash \{x_t\}$$

Once these eigenvectors $\widetilde{u_t}$ are obtained, we use absolute value of cosine similarity to measure the variation of the principal directions,

$$s_t = 1 - \left| \frac{\langle \tilde{u}_t, u \rangle}{\|\tilde{u}_t\| \|u\|} \right| \qquad \ldots (11)$$

This s_t can be considered as a score of outlierness, which indicates the anomaly of the target instance x_t [4].

VI. CONCLUSION

As per the above discussion it is necessary to detect credit card fraud to prevent loss of valuable personal and business assets. The main objective of this review is to construct a novel Outlier Detection System to detect credit card frauds. This system will work on online large scale data. Here we work on 20 attributes which will be reduced after applying PCA. On these reduced attribute set this technique will detect frauds with less memory and computation requirements. By using PCA it will be possible to detect only important attributes which contain major information. Due to those attributes such as credit history, purpose it will be possible to perform detection faster. In addition to this if certain transaction is a fraud it will get stored in database so that at next iteration this type of fraud will be detected by checking database.

Acknowledgements

I am very much thankful to my respected guide Prof. P. N. Kalavadekar for furnishing his valuable guidance for preparing this paper. I am also thankful to SRESCOE, Kopargaon and all my friends for their support.

REFERENCES

[1] Y Kou, C. Lu, S. Sinvongwattana and Y. Huang, Survey of Fraud Detection Techniques, *IEEE International Conference on* ***Networking,*** *Sensing & Control*, 21-23(3),2004.

[2] C. Aggarwal, *Outlier analysis* (Boston/Dordrecht/London, Kluwer Academic).

[3] T. Anh and S. Mägi, Principal Component Analysis, *Final Paper in Financial Pricing*, 2009.

[4] Y. Lee, Y. Yeh, and Y. Wang, Anomaly Detection via Online Oversampling Principal Component Analysis, *IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 7*, July 2013.

[5] M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, LOF: Identifying Density-Based Local Outliers, *Proc. ACM SIGMOD Int'l Conf. Management of Data*, 2000.

[6] V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection: A Survey, *ACM Computing Surveys, Vol. 41, No. 3*, 15:1-15:58, 2009.

[7] Lindsay I Smith, A tutorial on Principal Components Analysis,2002.

[8] W. Wang, X. Guan, and X. Zhang, A Novel Intrusion Detection Method Based on Principal Component Analysis in Computer Security, *Proc. Int'l Symp. Neural Networks*, 2004.

[9] Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, *IEEE Transactions On Knowledge And Data Engineering*.

[10] C.C. Aggarwal and P.S. Yu, Outlier Detection for High Dimensional Data, *Proc. ACM SIGMOD Internatinal Conf. Management of Data*, 2001.