Evaluation of Machine Learning Algorithms to Detect Credit Card Fraud

Sonal Mahajan, Dr. Shwetambari Chiwhane

Computer Department, NBN Sinhgad School of Engineering

Abstract

Credit card fraud detection is very serious issue nowadays. Every person from youth to aged requires a credit card. Credit card fraud is generally done during online transactions. Information regarding pin is obtained illegally. It is termed as shoulder surfing. Some other ways are card stealing, Buying audit cards, Information and web Traffic, etc. After obtaining information illegally online transactions are made. In many companies fraud credit cards are identified so that customers are not charged for unnecessary equipment. There is a vast need of credit card fraud detection. If proper amount of data is collected, credit card fraud can be detected using machine learning algorithms. In this paper, supervised and unsupervised machine learning algorithms have been applied to detect credit card frauds in a highly imbalanced dataset. It was found that unsupervised machine learning algorithms can handle the skewness and give best classifications result.

Keywords: *Credit card, fraud detection, machine learning, supervised learning, unsupervised learning.*

I. INTRODUCTION

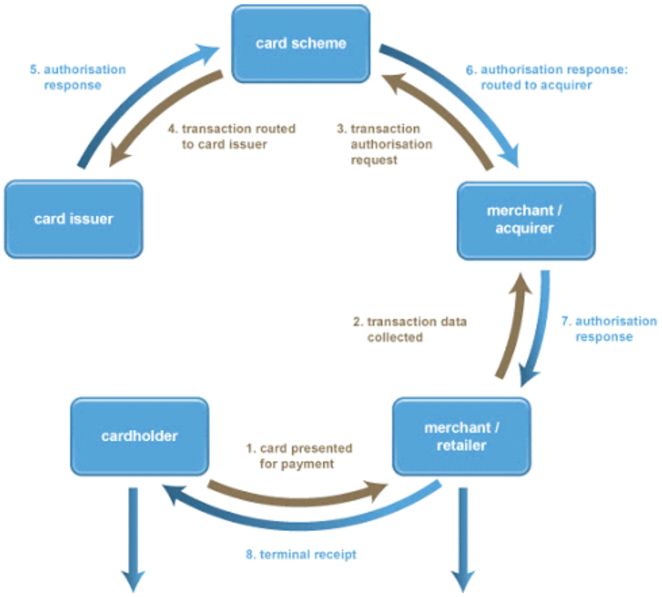
Fraud has been rising to a higher extend due to increasing use of online payments and social lifestyles.

Two best ways to avoid fraud are prevention of fraud and detection of fraud. Prevention is avoided by avoiding attack of any fraud. This can be done by acting as a barrier between the fraud and the system.

But if prevention is not successful then it may lead to fraud and then we need to focus on detection of fraud. It is the next step after prevention is failed.It is used to detect fraud. Credit card fraud happens generally by two ways: 1. Card present (CP) and 2. Card Not Present (CNP). We can understand by names itself that one type consist of card and other type user does not have card.Generally a recent survey shows that card not presence and card present are both common but card not present is usually performed by many. It is very common type of credit card fraud. According to various surveys conducted in October 2016, more than $31 trillion were generated worldwide by online payment systems in 2015, increasing 7.3% than 2014. Worldwide losses from credit card fraud have been rising to a greater extend in 2015, and will possibly reach $31 billion by 2020. However, there has been a vital increase in fraudulent transactions that affect the economy drastically. With increase in Fraud in any company, companion reputation also reduces. Issuer in the company has to resolve this problem. Many companies use a set of rules for employees working in their company. They have high security, passwords, codes, etc. In spite of all such measures if card information is stolen then card holder, issuing bank as well as merchant become victims of a credit card fraud, as it is one of them who has to bear the burden of fraud.

Machine learning can be termed as a solution to every generation which will reduce work and result in an better and efficient way to give output. It can work on large data sets of human being effectively. It is generally classified into two main types: Supervised and Unsupervised Learning. They are broadly described as follows. To detect fraud we require certain data sets. With the help of given data sets as input in an algorithm we can detect and prevent fraud with algorithm techniques provided. Supervised learning is based on classification of similarities and then working on them. In this generally input is known and output is also known. Only the method is to be decided i.e. After using method for a given input, known output should be obtained. In unsupervised learning, input is known and output is obtained based on method. It is most commonly seen that several machine learning algorithms are used to detect credit card fraud.

II. WORKING OF CARD SYSTEM


card scheme
5. authorisation response
6. authorisation response: routed to acquirer
4. transaction routed to card issuer
3. transaction authorisation request
card issuer
merchant / acquirer
2. transaction data collected
7. authorisation response
cardholder
1. card presented for payment
merchant / retailer
8. terminal receipt

Working of card system

III. METHODS TO DETECT FRAUD

1. Supervised Learning:

The majority of practical machine learning uses supervised learning method to detect fraud.

Supervised learning is an efficient way in which you can have output for a given input variable. If you have input variable as (x) and you will have an output variable (y). You use different algorithms for mapping function from valid input to validate output. Consider equation:

$$Y = f(X)$$

The main aim is to find mapping function for given input set to predict an output with given set of variables. You can do this by using any algorithm technique. Their are many algorithms mainly classified into supervised and unsupervised. Let us learn supervised algorithm in following section.

The name supervised algorithm simply tells us about the concept of teaching as a teacher. It consist of a set of data which is to be supervised and then output is calculated for the data. We know the answer, the algorithm itself repeats and makes predictions on the training data and is corrected by the teacher. Acknowledgement terminates when the algorithm is achieved at an valid level to perform. Supervised learning algorithm is mainly of two types as follows:

- Classification: It is a problem in which when the output variable is a type of category, such as “red” or “blue” or “disease” and “no disease”.
- Regression: It is a problem in which real value is termed as an output variable, such as “dollars” or “weight”.

• Random Forest - It is a method to learn classification, regression and other methods to assemble and operate set of data for construction of trees. These trees are decision trees and output is the mode of class or mean prediction of that class. In this a tree is constructed and evaluated to store and calculate data nodes are used.

• Neural Networks (NN) and Artificial Neural Network (ANN) - It is the most commonly used technique to detect credit card fraud in algorithms. It consist of group which is of neurons with interconnected nodes similar to links. These links appear to connect each other. These links have axon-synapse-dendrite connections. Strength of influencing node can be determined by weight of each link. It is used to detect effect of one node on other.

•Deep learning - It uses multiple layers to remove important features from the set of given raw data input. For example, consider in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.

• Support Vector Machine (SVM) - It is an analyzing tool used to analyze data which is used by classification and regression analysis. These are supervised learning models[7 8].

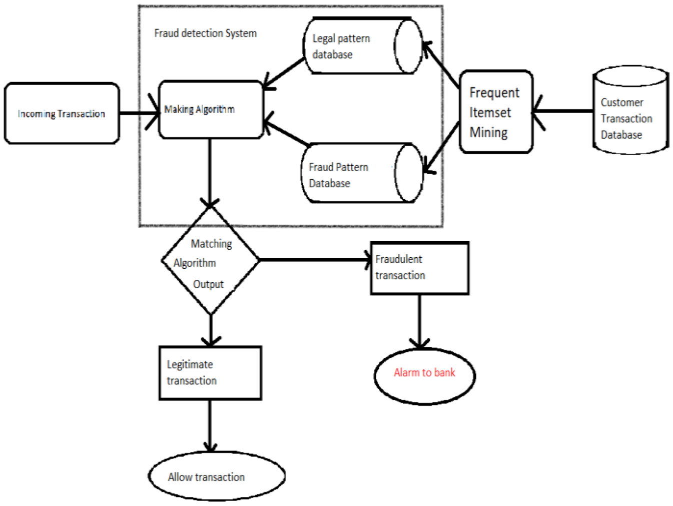
• Naive Bayes (NB) - The name itself states Bayes theorem. It uses Bayes theorem for classification of inputs. It is used for strong input set variables. It has independent attributes of data.

• Logistic Regression - It uses logical reasoning like probability which can be either one option or other. It finds chances to win or lose, alive or dead, rich or poor, healthy or sick. It can be best used for detection of images which may include different animals, fruits, etc.

• Extended Gradient Boosted Tree (XGBT) - It generally produces a tree which is used to determine if the given set of data is strong or weak. It is generally used for prediction which can be weak prediction of the form decision passed by decision tree.

• Quadratic Discriminant Analysis:It is closely related to linear discriminant analysis. It uses operator called Quadratic Discriminant Analysis and so has its name as QDA. It consist of measurements which are apparently scattered.

• K-Nearest Neighbours - n pattern recognition, used generally in which things related to parameters are not involved. It is beneficial for both classification and regression. It includes k training examples in feature space.



Fraud Detection System

2. Unsupervised Machine Learning:

In Unsupervised learning algorithms generally it does not have any output data. It performs its work on the input dataset only. It has given input and output is calculated on the given input. It mainly focuses on gaining the dataset by proper gaining of data or collecting them in group and then distributing them in an appropriate manner[9 10]. They are mainly categorized as unsupervised as they don't have any supervisor or a teacher on them. They have no guide for guidance. They even don't have correct output for input as an example to predict an algorithm. It is algorithm which is to be designed on its own by set of inputs without output. Clustering and association are two major types of unsupervised algorithms.

- Clustering: In this problem inherent group is to be discovered by grouping in the given input set, such as grouping material by need.
- Association: In this, it is considered as a rule in which we have to create certain rules which could give us valid output for a given set of input[6 11]. It generally refer to a large set of condition. Such as group of people has certain free ship based on age.
- Self-Organizing Maps (SOM): It is a type of ANN (Artificial Neural Network) under unsupervised learning to provide a small dimensional , different representation of input set of variables which are training samples and are called as map. It is used for dimensional reduction purpose.
- K-means:It is a method used to process a signal for analysis of clusters in mining of data. It can be termed as data mining process. It does partitions of given set of data into k clusters such that every data in under a cluster. Each cluster is selected in such a way that nearest cluster is calculated and categorized.
- Isolation Forest: It is used to detect anomaly and works on principle of isolating anomalies. It consist of observations and is generated by calculating mean. It is used to detect complex structures which are difficult to identify by a normal eye and so we use anomaly detection.
- The local outlier factor: It is used to find the deviations of a given data point to its nearest position in an area. It shares major techniques like "Core distance" and "Reachable distance". These are used for estimation of local density.

IV. RESULTS

After successfully detecting fraud and giving a set of input we need to examine the data if the output set is valid or not. For invalid input and fraud detection an alarm rises. Accuracy is considered as an important feature for alarm. If there is no accuracy or matching in data sets then alarm is invoked ad fraud is detected.

We create a confusion matrix for faster evaluation of fraud rate and alarm rate to catch fraud. We can acknowledge matrix by the confusion matrix as stated below:

		Expected	
		+ve	-ve
Reality	+ve	TP	FN
	-ve	FP	TN

In the given matrix as we see above we come to know that confusion matrix contains rows and columns in which row tells us about reality and columns tells us about expected output. TP stands for true positive which means that the transactions which occurred are result of transactions which can be considered as not fraud. They are real transactions. FP is False Positive ie. Transactions which occurred and their was an alarm but the alarm was false. It simply means that the transactions were not fraud but alarm occurred. FN is False Negative which consists of transactions that were fraud but their was no alarm. Fraud transactions are considered as true by mistake. TN is True Negative. It tells us that total transactions performed among which some where fraud and correctly identified as fraud with an alarm.To achieve a great rate to catch fraud and less alarm rate for false alarm is important to detect fraud.To find the true positive rate and false positive rate we use equations given below:

FPrate =	TP
	FN+TP

TPrate =	FP
	TN+FP

TP rate is the number of transactions that were correctly identified as fraud. FP rate is the number of transactions that were not correctly recognized as fraud. They were real but termed as fraud.

Accuracy can be termed as total transactions that can be correctly classified as fraud if fraud and successful if real. This can be done by a proper error control method. To limit the error and calculate accuracy we use following two equations:

Accuracy =	TN+FP
	TP + TN +TP

Error Control =	FN+FP
	FP +FN+TN +TP

V. CONCLUSION AND FUTURE SCOPE

Support Vector Algorithm is an efficient method to detect credit card fraud. By using above algorithms we can reduce number of fraud occurring in future and it will result in secure transactions of credits. It will also be beneficial for future generation. People will fill secure to invest and trust on banks and credit card system. More money will be available in banks and can be used by people for loans and other purposes. These are most flexible techniques for detection of fraud.By using this techniques higher accuracy can be determined and larger set of operations can be done easily. It is also useful for handling complex and large data. All this techniques can be used effectively in future in low cost to provide reliable and satisfactory output. It will also reduce the number of errors accounting in this techno world.

REFERENCE

[1] Low and Slow Is How the Credit Card Fraudsters Roll:https://www.threatmetrix.com/digital-identity-blog/fraudprevention/low-and-slow-is-how-the-credit-card-fraudsters-roll/

[2] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi, ”Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, Aug. 2018.

[3] L. Zheng, G. Liu, C. Yan and C. Jiang, ”Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity,” in IEEE Transactions on Computational Social Systems, vol. 5, no. 3, pp. 796806, Sept. 2018.

[4] Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), “Credit card fraud and detection techniques: a review”, Banks and Bank Systems, Volume 4, Issue 2, 2009 .

[5] Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, “A review of Fraud Detection Techniques: Credit Card”, International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012 .

[6] Vladimir Zaslavsky and Anna Strizhak,” credit card fraud detection using selforganizing maps”, information & security. An International Journal, Vol.18,2006.

[7] Shwetambari Kharabe, C. Nalini,” Robust ROI Localization Based Finger Vein Authentication Using Adaptive Thresholding Extraction with Deep Learning Technique”, Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 07-Special Issue, 2018.

[8] Shwetambari Kharabe, C. Nalini,” Using Adaptive Thresholding Extraction - Robust ROI Localization Based Finger Vein Authentication”, Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 13-Special Issue, 2018.

[9] Shwetambari Kharabe, C. Nalini,” Evaluation of Finger vein Identification Process”, International Journal of Engineering and Advanced Technology (IJEAT), International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S, August 2019.

[10] Udayan Birajdar, Sanket Gadhave, Shreyas Chikodikar, Shubham Dadhich, Shwetambari Chiwhane, “Detection and Classification of Diabetic Retinopathy Using AlexNet Architecture of Convolutional Neural Networks”, Proceeding of International Conference on Computational Science and Application, online 05 January 2020, pp 245-253.

[11] Dr. C. Nalini, Shwetambari Kharabe, Sangeetha S,” Efficient Notes Generation through Information Extraction”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S2, August 2019.

[12] Shwetambari Kharabe, C. Nalini , R. Velvizhi,” Application for 3D Interface using Augmented Reality”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-6S2,August 2019.