An Efficient Study of Fraud Detection System Using Ml Techniques

Chapter · January 2020

DOI: 10.1007/978-981-15-3284-9_8

3 authors:

s. Josephine Isabella
4 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Sujatha Srinivasan
SRM Institute for Training and Development, Chennai, India
31 PUBLICATIONS **306** CITATIONS
SEE PROFILE

Suseendran G.
VELS INSTITUTE OF SCIENCE, TECHNOLOGY & ADVANCED STUDIES (VISTAS), CHE…
149 PUBLICATIONS **1,038** CITATIONS
SEE PROFILE

An Efficient Study of Fraud Detection System Using Ml Techniques

S. Josephine Isabella, Sujatha Srinivasan, and G. Suseendran

Abstract The growing world has the transactions of finance mostly done by the transfer of amount through the cashless payments over the Internet. This growth of transactions led to the large amount of data which resulted in the creation of big data. The day-by-day transactions increase continuously which explored as big data with high speed, beyond the limit of transactions and variety. The fraudsters can also use anything to affect the systematic working of current fraud detection system (FDS). So, there is a challenge to improve the present FDS with maximum possible accuracy to fulfill the need of FDS. When the payment is made by using the credit cards, there is chance of misusing the credit cards by the fraudsters. Now, it is essential to find the system that detects the fraudulent transactions as a real-world challenge for FDS and report them to the corresponding people/organization to reduce the fraudulent rate to a minimal one. This paper gives an efficient study of FDS for credit cards by using the machine learning (ML) techniques such as support vector machine, naïve Bayes, K-nearest neighbor, random forest, decision tree, OneR, AdaBoost. These machine learning techniques evaluate a dataset and produce the performance metrics to find the accuracy of each one. This study finally reported that the random forest classifier outperforms among all the other techniques.

Keywords FDS · Naïve Bayes · Random forest · SVM · Decision tree · OneR

S. Josephine Isabella (✉)
Research Scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, (VISTAS), Chennai, Tamil Nadu, 600117, India
e-mail: josephineisabella@yahoo.co.in

S. Srinivasan
Associate Professor, Department of Computer Science and Applications, SRM Institute for Training and Development, Chennai, Tamil Nadu 600032, India
e-mail: ashoksuja08@gmail.com

G. Suseendran
Assistant Professor, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies, (VISTAS), Chennai, Tamil Nadu, 600117, India
e-mail: suseendar_1234@yahoo.co.in

S.-L. Peng et al. (eds.), *Intelligent Computing and Innovation on Data Science*,
Lecture Notes in Networks and Systems 118,
https://doi.org/10.1007/978-981-15-3284-9_8

1 Introduction

The new arrival of innovative technologies gives an opening to the Internet and cashless transactions which have emerged as easier. However, for online transactions, we no longer want to be in a view found in a sure location where the transaction happens, making it prone to fraudulent one. There are many ways in which the people can profess to be the other user and create a transaction as fraudulent. If a transaction is fraudulent or no longer available, it could be decided by studying previous transactions and evaluating them with the modern one. If the distinct in nature of previous transaction and the modern transaction is big, there is a possibility that the modern-day transaction is a fraudulent transaction [1]. This paper discusses an effective study about the machine learning techniques that detect the fraudulent transactions with the help of evaluation metrics in an effective way. Section 1 gives the introduction. A common study to understand the fraud detection system (FDS) is discussed in Sect. 2. Section 3 reviews the related literatures in FDS. Sect. 4 gives the experimental studies. The evaluations of various machine learning techniques are detected in Sect. 5. Section 6 gives the results and discussion part. Finally, the conclusion is given in Sect. 7.

2 An Understanding of FDS

Without using cash, the products can be sold and transferred through various payments by simply using a card that is given by the financial sectors and the bank called credit cards. The fraudsters use these cards illegally, or not having the permission of cardholders is referred to as credit card fraud [2]. The method used to find and identify the fraudulent transactions when the transactions have entered into the system and make intimation to a system administrator is called FDS. Previously, these transactions were obtained by using fraud detection sampling techniques, but it was time consuming. Nowadays, machine learning plays a major role in automated system [3]. The continuous increase of usage of credit card transactions and evolving the concept of CNP (card-not-present) in payment transactions that generate the misbehavior of the illegitimate people who counterfeit as others. There is a need to create an automated FDS for credit issuers [4]. So, there is a chance to apply the machine learning techniques to find the solution to the fraud detection system in a functional way [3].

3 Review of Literature

The study given by authors like Shen et al., investigated that the efficiency of classification models is tested against fraud detection and also produced a framework to the fraud detection in credit card to reduce the risk [5] at banks. Whitrow et al., revealed a study of fraud detection at transaction and account level of two banks, A and B, by using the transaction aggregation [6]. In this proposed study, the self-organizing map neural network (SOMNN) technique and transactional rules are used to create a decision model called credit card fraud watch (CCFW) along the existing banking software and are applied to the real banking dataset and used to solve the problem of fraudulent transaction by the optimal classification of each transaction [7].

The authors, E. Duman and Y. Sahin, designed a model for fraud finding and discussed that SVM models produced better results in the training dataset mode, while the decision tree-based models performed well in the testing mode. This model can be utilized by the financial institutions to predict the fraudulent transaction. [8]. This study implemented a linear Fisher discriminant analysis on fraud detection in credit cards for calculating a weighted average to find out the transactions as profitable and prevented loss of millions of dollars of real-time banking transactions [9].

Awoyemi et al. [10] concluded that there is a need to develop a better sampling approach to handle the highly imbalanced credit card dataset using meta-classifiers. This study made a comparison of random forest and logistic regression with sample dataset (preprocessed with PCA and without PCA values). This comparison evaluated through the R language resulted that Random forest without PCA and a K value of 3 having the accuracy as 99.77% by using the confusion matrix [11]. This study is designed to build four classification models, namely logistic regression, SVM, decision tree and random forest with the training data of 70 and 30% testing data of European card holders from ULB Machine Learning Group. Random forest is found as the best classifier among all [12]. John et al. made an effective study of feature selection on two imbalanced datasets as ranking by the use of correlation coefficient and evaluated using MATLAB IDE with the four classifier techniques, namely naive Bayes, support vector machine, decision tree and NNBRF and applied to the datasets of Taiwan and European banks. The results showed that the decision trees were performed to produce the better result of classification [13].

Rajora et al. made a study of machine learning classification techniques as well as ensemble learning methods and evaluated an unbalanced dataset by using under sampling method with PCA values as balanced. The outcomes showed that the gradient boosting regression tree had the better accuracy among all the classifiers based on dataset 'without time' feature [14]. Authors like patil et al. evaluated the random forest, logistic regression and decision tree classifiers and applied on the credit card fraud-German dataset and results showed that random forest tree made accuracy as high but had the limitation of over fitting of decision tree [15]. K. R. Seeja and Masoumeh Zareapoor revealed a model named FraudMiner for fraud detection and analyzed the results of classification models. The FraudMiner model was applied to one lakh transactions. This proposed model produced the performance evaluation as fraud detection rate was high. The evaluation was done by applying the BCR and MCC to the FraudMiner model [16].

In this study, the authors reviewed various methods to find the solution to the fraud detection systems. They discussed hidden Markov model (HMM), CNN and ANN methods and proposed a model with autoencoder neural network model [1].

4 Experimental Studies

4.1 Dataset and Preprocessing of Data

The German credit fraud dataset is the famous dataset taken from kaggle.com with 1000 instances and 20 attributes. Preprocessing is essential before we evaluate the values in the dataset. The proposed model gives the accuracy improvement based on the features that have been selected as salient features. In this study, we use the German credit card dataset as sample dataset. The model has been trained with 70% of instances and tested with 30% of instances having 20 attributes [17].

4.2 Evaluation Metrics

There are some metrics of evaluation available to find the achievement measures of the classification models.

The various metrics for evaluation are given as follows [10, 20]:

$$\text{Accuracy} = (\text{TN} + \text{TP})/(\text{TP} + \text{FP} + \text{FN} + \text{TN}) \tag{1}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{2}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{3}$$

Based on the evaluation of these metrics, the confusion matrix is formed.

5 Evaluation of Ml Techniques

5.1 Naïve Bayes

Based on some assumption, the outcome is affected by the independent factor that is called as ‘Naive.’ It predicts a class of future incoming data values with known target values as training data. It finds the probability by using the formula [16, 18].

5.2 KNN

This algorithm predicts data value based on a relative position to other data values. It is a clustering algorithm used to find the unknown feature of a testing data by using the Euclidean distance [18]. This is an instance-based algorithm which keeps all the instances and classifies the similar instances having the nearest values. The existing instances find the new nearest instances by using distance evaluation such as Euclidean distance [16].

5.3 Random Forest

The Random Forest classifier generates the connected decision tree classifiers randomly. If the input is having the training data, then it will make the rules which are helpful to predict the results through the decision tree forests [18]. This technique generates a decision tree having the concept as each tree is a weak learner and the tree having maximum votes are the strong learners, and it categorizes the new instances to the class that has the maximum votes [16].

5.4 SVM

For classification problems, SVM is used to categorize the values or data points by the best fitting method. Support vector machine plots the line that denotes the training values on a plane to detect the categorization of data. The classification problems and regression model problems use this technique in an efficient way to find the solution [18].

5.5 Decision Tree (J48)

J48 is a decision tree model and an implemented form of C4.5 technique in Java. This is an ID3 decision tree algorithms extended version. Working on the different values of an existing input, the average value of new class can be calculated. The different features are represented in the tree as internal nodes. The end value of the dependent data is found by the end node. The root node gives the decision.

5.6 OneR

The frequency table has target value for each predictor for creating a predictor's rule called one rule that selects the rule that has the minimum total error.

5.7 AdaBoost

This algorithm is a classification ensemble method. This algorithm is used to improve the performance of any algorithm. When any algorithm combines with this technique, then it converts the weak learners to the strong one [19].

6 Results and Discussions

The evaluation of machine learning techniques produces the results of various measures such as the rate of true positive, precision and are related to find the fraudulent transactions in an efficient way. These measures are observed and placed in Table 1.

Obviously, all the ML techniques produced true positive greater than 80%. The random forest algorithm has the highest rate of true positive as 92%. The remaining techniques have less than that of 92%. The SVM and OneR techniques having the same true positive rate 87% are slightly higher than naïve Bayes. KNN has the lowest rate (81%) of true positive. The Recall value of random forest attained at the maximum of 0.917 and KNN has the lowest value 0.810 of Recall. SVM has the recall value as 87.1% and is slightly higher than that of OneR and naïve Bayes methods. The transactions which are correctly classified as genuine or fraudulent are usually termed as precision. From the evaluated results, naïve Bayes classifier has the most prominent precision value as 80% and OneR method has the lowest value of 71.2%. The next highest precision value obtained by KNN is 79.4%. But the KNN algorithm has the lowest rate (19%) of false positive. This shows that this algorithm

Table 1 Results of classification measures using various ML techniques

Ml technique	TPR (%)	FPR (%)	F-measure	MCC	Recall	Precision	Acc
Naïve Bayes	86	50	83.10	0.385	0.864	0.800	75.4
KNN	81	19	80.19	0.324	0.810	0.794	72.0
Random forest	92	59	84.47	0.386	0.917	0.783	76.4
SVM	87	53	83.05	0.371	0.871	0.793	75.1
J48	84	61	95.00	0.250	0.840	0.763	70.5
OneR	87	82	78.17	0.061	0.867	0.712	66.1
AdaBoost	88	73	80.10	0.180	0.877	0.737	69.5

handles the dataset in a better way than other classifiers. The remaining techniques having FPR greater than 19% observed from the evaluation.

The correctly classified instances of incoming data categorized after evaluating the various machine learning techniques are shown. The accuracy results are represented in Fig. 1. From the graph, random forest algorithm has the most accurate value as 76.4%. The least accuracy is produced by OneR method. The naïve Bayes classification and SVM have more or less the same accuracy with the difference of 0.3%. J48 decision tree algorithm classifies the data with the accuracy rate of 70.5%. The AdaBoost algorithm detects 69.5% of accuracy, but is greater than that of OneR method which has the accuracy of 66.1%.

Random forest technique has the highest (84.5%) F-measure value. SVM and naïve Bayes have the same value, 83%. Similarly, KNN and AdaBoost have the same value (80.1%) for the F-measure. This observation is visualized in Fig. 2.

Matthews correlation coefficient (MCC) [7, 17] has been calculated for various machine learning models. The Matthews correlation coefficient must be in the range

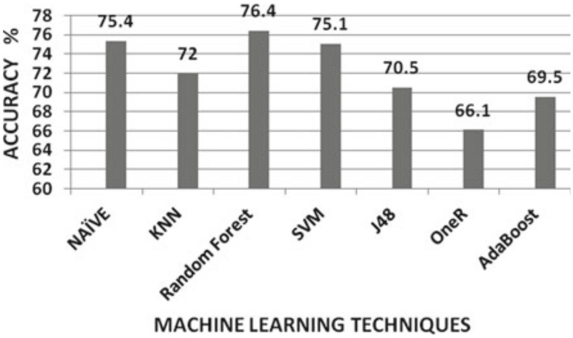

Fig. 1 Accuracy of FDS using ML techniques

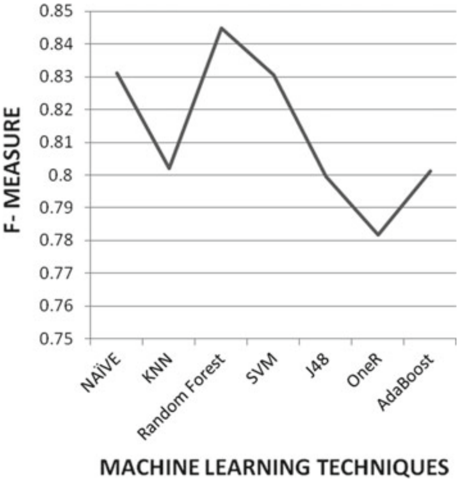

Fig. 2 Comparative analysis of F-measure

of +1 to −1. All our evaluated techniques resulted in this range of values and are efficient to fit in the model.

7 Conclusion

Usually, the available fraud detection methods find the fraudulent transaction after they have happened. There will be a chance to occur fraudulent transaction out of numerous transactions. Even though the occurrence of fraud is at minimal rate against large number of transactions, it is a commitment to invent a technique for detecting the fraudulent cases before the transaction has been completed. This study made an effort to evaluate the sample dataset with different machine learning techniques and resulted that among all the techniques random forest technique produces better performance in most of the cases. The above study showed that the machine learning techniques are capable of handling the fraudulent cases in an efficient manner. But there is a limitation occurred that how their performance will be found when the total number of transactions will be increased to some extreme level, i.e., how they are scalable. This experimental study gives a pathway to find an efficient highly scalable machine learning technique. There is a need to create a framework that handles the big data in a smooth way to find the fraudulent transactions at a minimal rate in the field of fraud detection system as future work.

References

1. Manek H (2019) Title : review on various methods for fraud transaction to secure your paper as per UGC guidelines we are providing a electronic bar code, Nov 2018
2. Chaudhary K, Yadav J, Mallick B (2012) A review of fraud detection techniques: credit card. Int J Comput Appl 45(1):975–8887
3. Abdallah A, Maarof MA, Zainal A (2016) Fraud detection system: a survey. J Netw Comput Appl 68:90–113
4. Van Vlasselaer V et al (2015) APATE: a novel approach for automated credit card transaction fraud detection using network-based extensions. Decis Support Syst 75:38–48
5. Aihua S, Rencheng T, Yaochen D (2007) Application of classification models on credit card fraud detection. In: Proceedings-ICSSSM'07 2007 International Conference Service System Service Management, no. 1997, 2007, pp 2–5
6. Whitrow C, Hand DJ, Juszczak P, Weston D, Adams NM (2009) Transaction aggregation as a strategy for credit card fraud detection. Data Min. Knowl. Discov. 18(1):30–55
7. Ogwueleka FN (2011) Vol_6(3)_311-322_Ogwueleka.pdf. 6(3):311–322
8. Sahin Y, Duman E (2011) Detecting credit card fraud by decision trees and support vector machines. Int Multiconference Eng Comput Sci I:6
9. Mahmoudi N, Duman E (2015) Detecting credit card fraud by modified fisher discriminant analysis. Exp Syst Appl 42(5):2510–2516
10. Awoyemi JO, Adetunmbi AO, Oluwadare SA (2017) Credit card fraud detection using machine learning techniques: a comparative analysis. In: Proceedings of the IEEE International Conference Computing Networking Informatics, ICCNI 2017, 2017, vol 2017-Jan, pp 1–9

11. Data T (2017) A comparison of machine learning techniques for credit card fraud detection, pp 1–9, 2017
12. Navanshu Khare SYS (2018) Credit card fraud detection using machine learning models and collating machine learning models. J Telecommun Electron Comput Eng 10(1–4):23–27
13. John OA, Adebayo A, Samuel O (2018) Effect of feature ranking on the detection of credit card fraud: comparative evaluation of four techniques. i-manager's J Pattern Recogn 5(3):10
14. Rajora S et al (2019) A comparative study of machine learning techniques for credit card fraud detection based on time variance. In: Proceedings 2018 IEEE Symposium Series Computational Intelligent SSCI 2018, no Nov, pp 1958–1963, 2019
15. Patil S, Nemade V, Soni PK (2018) Predictive modelling for credit card fraud detection using data analytics. Procedia Comput Sci 132:385–395
16. Seeja KR, Zareapoor M, FraudMiner: a novel credit card fraud detection model based on frequent itemset mining. Sci World J, vol 2014, 2014
17. Correa Bahnsen A, Aouada D, Stojanovic A, Ottersten B (2016) Feature engineering strategies for credit card fraud detection. Expert Syst Appl 51:134–142
18. Banerjee R, Bourla G, Chen S, Purohit S, Battipaglia J (2018) Comparative analysis of machine learning algorithms through credit card fraud detection, pp 1–10
19. Sun Y, Wong AKC, Wang Y (2010) Parameter inference of cost-sensitive boosting algorithms, pp 21–30
20. Jain Y, NamrataTiwari SD, Jain S (2019) A comparative analysis of various credit card fraud detection techniques. Int J Recent Technol Eng 7(5S2):402–407