A Random Forest Fraud Transaction Detection System using Data Mining Techniques

S. BAKIYALAKSHMI

Assistant Professor, CSE DEPARTMENT

Aalim Muhammed Salegh College of Engineering

bakiyalakshmi.s@aalimec.ac.in

R. PRIYADHARSHINI

Final Year Student, CSE DEPARTMENT

Aalim Muhammed Salegh College of Engineering

priyadharshini482000@gmail.com

***Abstract*—Every day, news of financial statement fraud is adversely affecting the economy worldwide. Considering the influence of the loss incurred due to fraud, effective measures and methods should be employed for prevention and detection of financial statement fraud. I integrate scholastic writing identified with fake monetary revealing with double purposes: (1) to more likely comprehend the nature and degree of the current writing on monetary announcing misrepresentation, and (2) to feature regions where there is need for future examination. I survey distributions in bookkeeping and related disciplines including criminal science, morals, finance, authoritative conduct acknowledged for distribution. The implementation of data mining techniques for fraud detection follows the traditional information flow of data mining, which begins with feature selection followed by representation, data collection and management, pre -processing, data mining, post-processing, and performanceevaluation**

***Keywords*—Fraud detection, Data mining, Online Transaction**

I.INTRODUCTION

With the Internet technology into all aspects of people's lives, whether in the field of trade or finance, the degree of informatization and virtualization is deepening. In recent years, with the rapid expansion of the frequency and scale of online transactions, more and more people use the Internet to shop. At the same time, the amount of online transactions is also increasing. All these are fertile soil for the breeding of transaction fraud. Fraud criminals often use various channels to steal user information and transfer a large amount of money in the shortest time, causing a lot of property losses to users and banks. Transaction fraud is often an abnormal event hidden in many daily transactions in the financial field. In order to avoid huge labor costs, it is a trend to detect abnormal trading behaviour by means of machine learning and data mining. The core detection algorithms of the system are mainly based on classification [1-5]. In order to face a large amount of data, data mining related processing methods are introduced into transaction fraud task [6-11]. In this paper, we establish a fraud detection system based on the classification model of random forest and the data processing related to feature engineering.

A. RELATED WORK

Data mining is a process of exploring information hidden in a large amount of data through algorithms. Through the cleaning, correction, extraction, selection, and summary of a large number of data features, the hidden knowledge behind the data is obtained. For our problem, it is to extract the difference information between the behaviour patterns of real users and fraud behaviour patterns in the data, so as to help the subsequent classification model better achieve the purpose of detecting fraudulent transactions. In [3], Fang et al. Proposed a framework for fraud detection based on CNN to capture the inherent patterns of fraud learned from the marked data. Wang [5] proposed a data mining algorithm based on UCI public dataset. The research of work [8] mainly focuses on the most important feature engineering part of data mining. Researchers extend the transaction aggregation strategy and propose to create a new set of features by using von Mises distribution on the basis of analysing the periodic behaviour of transaction time.

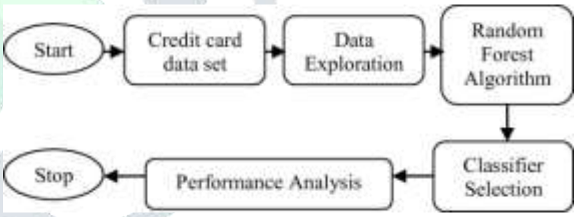


Figure 1. Fraud Detection using Random forest Algorithm

However, these methods have some defects. In reference [5], due to the limited amount of data, it is difficult to establish a robust and accurate fraud detection system for practical scenarios. Besides, most of these techniques are absence of manual grouping. Manual characterization can compensate for the lack of programmed grouping and further stay away from the danger of misrepresentation.

B. Our Contribution

Referring to the boost-based fraud transaction model [9], this paper proposes a semi-automatic fraud transaction detection system based on random forest (as shown in Figure 1). Among them, the automation part is a fraud transaction risk detection model based on random forest, and the core of the other half is an expert reviewer. If the output risk of the risk detection model is higher than the threshold value, it will be regarded as a high-risk transaction and transferred to the expert reviewer. The expert review will combine the expertise and the information provided by the risk detection model to make future judgement.

The data of training risk detection model is IEEE-CIS data set. The data set contains more than 1 million samples, and each sample contains more than 400 characteristic variables, including financial characteristics and nonfinancial characteristics. The richness of data will significantly improve the accuracy of the fraud detection model and avoid the problem of over fitting. Of course, complex data need feature engineering to do some processing.

We first clean the data to eliminate some outliers and missing data. Furthermore, the data are transformed and the statistical data such as maximum, mean, and standard deviation are extracted. Then, Recursive feature elimination (RFECV) is used to eliminate some unimportant features. Finally, we implement a classifier based on random forest to detect transaction fraud risk.
In order to show its superiority, we compare it with support vector [12-21] machine and logistic regression. The experimental results show that our model has achieved good results in accuracy and ROC AUC score.

The rest of this paper is arranged as follows. The second section introduces the characteristic engineering of financial data and non-financial data. The third section introduces our Semi automatic fraud transaction detection system based on random forest. In the fourth part, the performance of this algorithm is compared with other classical machine learning models by accuracy and AUC ROC score. Finally, the fifth part summarizes this paper

II. FEATURE ENGINEERING

The IEEE CIS fraud data set is provided by Vesta, which is the pioneer of e-commerce payment solutions. The dataset is divided into two groups of tables: transaction table and identity table.

These two kinds of tables are connected by key transaction ID, but not all transactions have identity information. The data has been labelled as two categories, i.e. is fraud = 0 or 1, but other characteristic information is still messy. Therefore, we first conduct data mining on them, and then combine them to form the final training data. In order to better handle data, we can study the two tables separately. Unfortunately, there are hardly any credit card datasets publicly available for study due to the private nature of financial transactions. Lopez-Rojas et al. (2016) in their paper PaySim: A financial mobile money simulator for fraud detection propose a simulation tool called PaySim to generate similar transactions based on their original mobile money transaction dataset. The synthetic dataset is available on Kaggle.com.

TABLE 1-Transaction table

Name	Description	Type
Transaction ID	ID of transaction	ID
isFraud	binary target	categorical
Transaction DT	transaction date	time
TransactionAMT	transaction amount	numerical
card1-card6	card	categorical
addr1-addr6	address	categorical
M1-M9	anonymous features	categorical
P_email domain	Purchaser email doma	categorical
R_email domain	receiver email domain	categorical
dist1-dist2	country distance	numerical
C1-C14	anonymous features	numerical
D1-D15	anonymous features	numerical
V1-V339	anonymous features	numerical

Table2-Identification table

Name	Description	Type
TransactionID	ID of transaction	ID
DeviceType	device type	categorical
DeviceInfo	Device Information	categorical
id01-id11	identification data	numerical
id12-id38	Identification data	categorical

A. Transaction table

The transaction table has 394 characteristic variables, including 22 classification features and 372 numerical features. Most digital features are anonymous with fixed prefixes. To give a specific and clear description, we summarize these variables in Table 1. Transaction refers to the transaction date and time, which can be parsed into precise time information, such as year, month, day, week, etc. Transaction MT refers to the amount of transaction payment in U.S. dollars. A small part of the amount with irregular decimal, may represent the transaction for remittance calculation.

B. Identification table

The distinguishing proof table contains 41 features, including personality data, network association data related with exchanges (IP, ISP, specialist, and so on) and conduct data. At the point when the two tables are then handled independently, they are joined on the value-based key to produce another table. It likewise records conduct fingerprints, for example, account login time and login disappointment time, account span remaining on the page, etc. Our data mining methods mainly include data cleaning, missing value filling, data transformation and feature extraction. In the data cleansing section, we deleted columns with a large percentage of Nan values (missing values).

For example, if more than 90% of the value in a feature column is Nan, the column is deleted. For data that is not very defective, we'll fill it with - 1000 (a specific value that doesn't appear in the data). For the time code, we will also convert it into more accurate time information for better use. In addition, we also generate many descriptive statistical features, such as the mean and extreme value of numerical characteristics such as transaction amount, billing address and mailing address. Finally, a large part of the digital features has correlation. It can greatly improve the efficiency of data fitting and improve the performance of data classification. Therefore, in our work, recursive feature elimination with cross validation (RFECV) is used to eliminate each feature iteratively.

III.RANDOM FOREST FRAUD DETECTION MODEL

In this part, we will introduce the Random Forest risk detection model.

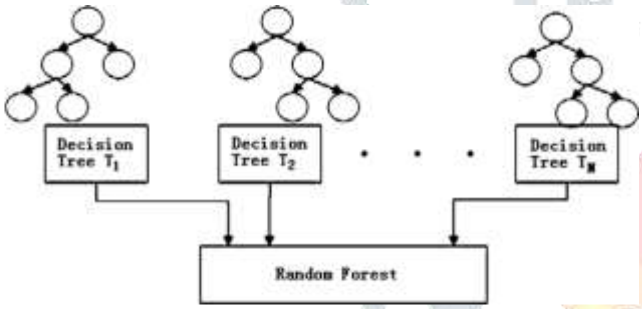


Figure 2. Part of a decision tree in a random forest

Random forest is a classifier with various decision trees. It has adaptable model and quick preparing. It can address the characterization blunder brought about by the very unequal information of misrepresentation exchange identification. Each tree is produced by an irregular vector of free testing, and each tree votes to track down the most famous classification to arrange the information. Irregular backwoods has both example haphazardness and trademark arbitrariness, and its speculation execution is prevalent. Simultaneously, arbitrary woodland has great handling capacity for high-dimensional informational collections, which is entirely appropriate for IEEE CIS informational indexes. It can handle countless information sources and decide the main attributes. In this way, further element mining is completed on the information extricated by RFECV.

IV.EXPERIMENTS

In order to show the superiority of the model, we compared the accuracy and AUC ROC score with other models. AUC ROC score is actually the area under the receiver operating characteristic curve, which is created by drawing the relationship between true positive rate (TPR) and false positive rate (FPR) under different threshold settings. The formulas for TPR and FPR are defined as follows:

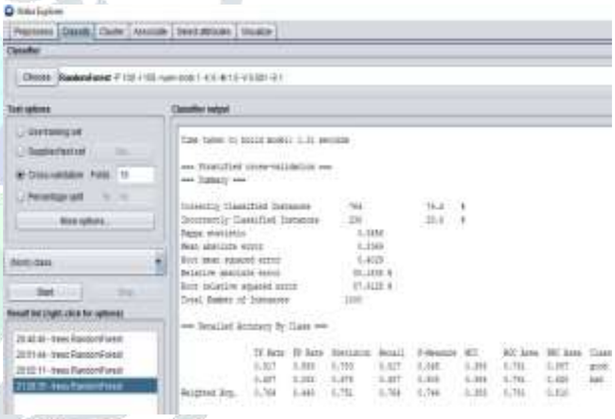
$$\underset{\textit{also called sensitivity/recall/hit rate}}{\text{True Positive Rate (TPR)}} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP + FN}}$$

$$\underset{\textit{also called fall out}}{\text{False Positive Rate (FPR)}} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP + TN}}$$

TP was the true positive prediction, FN was the false negative prediction, FP was the false positive prediction, and TN was the true negative prediction. The configuration matrix corresponding to the model can also be obtained (Fig. 2). As can be seen from table 3, the random forest model is superior to the other two models in terms of AUC ROC score and accuracy

A. DATASET

In this experiment, we used credit dataset based on the credit card data implementation is done using WEKA tool for analysing the Accuracy using different data mining Techniques that understood against the benchmark database.



```
Time taken to build model: 1.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        764               76.4    %
Incorrectly Classified Instances      236               23.6    %
Kappa statistic                         [illegible]
Mean absolute error                     [illegible]
Root mean squared error                 [illegible]
Relative absolute error                 [illegible]
Root relative squared error             [illegible]
Total Number of Instances            1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 [illegible]                                                                   good
                 [illegible]                                                                   bad
Weighted Avg.    0.764    [illegible]
```

TABLE 3 Performance of different models

Models	ROC Area	Accuracy
LMT	0.792	75.1
SMO	0.671	75.1
Logistic Regression	0.785	75.2
Random Forest	**0.791**	**76.4**

V. CONCLUSION

In this paper, a self-loader extortion recognition framework dependent on arbitrary woods is proposed, and its adequacy is checked on IEEE-CIS informational index. The subsequent segment presents the handling of informational indexes, including information cleaning, highlight choice and element designing. The third section introduces the transaction fraud risk detection model based on random forest. In the fourth part, the performance of the algorithm is compared with other two classical machine learning models by accuracy and AUC ROC score.

References

[1] Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Cluster based outlier detection. Annals of Operations Research, 168(1), 151-168.

[2] Minastireanu, E. A., & Mesnita, G. (2019). Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur, 2019.

[3] Fang, Y., Zhang, Y., & Huang, C. Credit Card Fraud Detection Based on Machine Learning.

[4] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261- 270).

[5] Wang, M., Yu, J., & Ji, Z. (2018). Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model.

[6] Dhingra, S. (2019). Comparative Analysis of algorithms for Credit Card Fraud Detection using Data Mining: A Review. Journal of Advanced Database Management & Systems, 6(2), 12-17.

[7] Minastireanu, E. A., & Mesnita, G. (2019). Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur, 2019.

[8] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142..

[9] Zhang, Y. , Tong, J. , Wang, Z. , & Gao, F. . (2020). Customer Transaction Fraud Detection Using Xgboost Model. 2020 International Conference on Computer Engineering and Application (ICCEA).

[10] Bhusari, V., & Patil, S. (2016). Study of hidden markov model in credit card fraudulent detection. In 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave) (pp. 1-4). IEEE.

[11] Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. Decision Support Systems, 95, 91-101.

[12] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[13] Suykens J A K , Vandewalle J . Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters, 1999, 9(3):293-300.

[14] Furey T S , Cristianini N , Duffy N , et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10):906-14.

[15] Tong S , Koller D . Support vector machine active learning with applications to text classification^]// JMLR.org, 2002:999-1006.

[16] Furey, T, S, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. BIOINFORMATICS -OXFORD-, 2000.

[17] Tong, Simon, Koller, et al. Support Vector Machine Active Learning with Applications to Text Classification.^]. Journal of Machine Learning Research, 2002.

[18] NGUYEN DUNG DUC MATSUMOTO KAZUNORI TAKISHIMA YASUHIRO. Re-learning method for support vector machine[J]. 2009.

[19] Heumann, Benjamin, W. An Object-Based Classification of Mangroves Using a Hybrid Decision Tree-Support Vector Machine Approach.[J]. Remote Sensing, 2011.

[20] Dong-Xiao N , Yong-Li W , Xiao-Yong M A . Optimization of support vector machine power load forecasting model based on data mining and Lyapunov exponents[J]. Journal of Central South University of Technology, 2010, 17(002):406-412.

[21] Keerthi S S . Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms[J]. IEEE Transactions on Neural Networks, 2002, 13(5):1225.

[22] Malinowski, Mariusz, Jasinski, et al. Simple Direct Power Control of Three-Phase PWM Rectifier Using Space-Vector Modulation (DPCSVM).[J]. IEEE Transactions on Industrial Electronics, 2004

.