Proceedings of the 2017 International Conference on Machine Learning and Cybernetics, Ningbo, China, 9-12 July, 2017

A HYBRID SEMI-SUPERVISED APPROACH FOR FINANCIAL FRAUD DETECTION

JIN-MIAO LIU, JIANG TIAN[1], ZHU-XI CAI, YUE ZHOU, REN-HUA LUO, RAN-RAN WANG[2]

[1] Department of Information Technology, China Everbright Bank, Beijing, China
[2] Department of Finance, Beijing Institute of Big Data Research, Beijing, China
E-MAIL: liujm@cebbank.com, tianjiang@gmail.com, zhuxic@bibdr.org, yuez@bibdr.org, renhual@bibdr.org, ranranw@bibdr.org

Abstract:

In this paper, we create a semi-supervised methodology for financial fraud detection in bank wire transactions based on a clustering-based-isolation-forest (CBiForest) algorithm. To test this hybrid model, we experiment on wire transaction data of twelve months from China Everbright Bank. The result of abnormal users is proved to be reliable and outperforms other clustering algorithms. Furthermore, our model can be regarded as a huge improvement for traditional expert system in bank.

Keywords:

Fraud Detection; Isolation Forest; K-means Clustering; Semi-supervised Learning

1. Introduction

The latest report [1] shows that more than $181 billion has been lost due to bank fraud in the United States 2016. The number will keep increasing in the next decade which attracts more concern across bank industry. Among various fraud types, wire fraud is one of the most damaging and takes up to $50 billion being estimated by FBI, industry reports and fraud executive interviews.

Machine-learning techniques were introduced for fraud detection since 1990s and became widely used because of high volume data and timeliness requirement. Kokkinaki (1997) [2] proposed the idea of a similarity tree using decision tree logic which labels nodes with attribute names and edges with attribute values satisfying some condition. Bentley et al. (2000) [3] employed genetic algorithms in order to establish a logic rule with highest predictability to classify transactions into suspicious and non-suspicious classes. Bolton & Hand (2002) [4] developed peer group analysis and breakpoint analysis to identify behavioral fraud among accounts and individual basis. Dorronsoro et al. (1997) [5] designed an online fraud detection system based on neural network and Maes et al. (2002) [6] applied Bayesian networks in the credit card industry.

All the techniques mentioned above are supervised learning algorithms in which massive fraud samples are required to train the models or systems. However, banks in developing countries are lack of enough number of training samples in wire transaction fraud because of insufficient data collection and incomplete data storage. Expert system is employed by most banks as a traditional method to detect fraud which consists of rules made by bank workers based on experience and fraud cases. Nevertheless, expert system has several shortcomings, such as bad timeliness, high false positive rate and expensive maintenance costs.

Therefore, in this paper we propose a novel hybrid semi-supervised learning approach for bank fraud detection in wire transaction and experiment with 80 million transaction records data during 2016 from China Everbright Bank (CEB). The rest of the paper is organized as follow: Section 2 reviews some unsupervised learning algorithms in anomaly detection. We apply exploratory data analysis, feature engineering and model building in Section 3, 4, 5. Section 6 and Section 7 are discussion and final conclusion.

2. Algorithm

Here, we review two main unsupervised learning methods: Isolation Forest (iForest) and simple k-means clustering (SKM). These two algorithms have proved effective in anomaly detection during recent years and contribute to our hybrid approach which will also be introduced at the end of this part.

2.1. iForest and SKM

The theoretical process of iForest was proposed by Liu et al. (2008) [7,8] under the hypothesis that anomalies are data points which are few and different. Thus, iForest

978-1-5386-0408-3/17/$31.00 ©2017 IEEE

detects anomalies purely based on the concept of isolation without any distance or density measure which fundamentally different from other model-based methods.

Comparatively, SKM scheme [9] has been applied to classify transactions according to distance. This algorithm assigns initial points to two clusters by recursively partitioning and the smaller cluster is usually considered as fraudulent.

2.2. CBiForest

These two algorithms have their strength and weakness in the application of anomaly detection. The iForest algorithm is able to isolate a cluster of anomalies but short in detecting abnormal points from various groups simultaneously. In the contrast, SKM has the capability in handling anomalies with different patterns based on distance but might be trapped in local optimum and high false alarm.

In this situation, we develop a clustering-based-isolation-forest (CBiForest) algorithm for anomaly detection which combines the strength of iForest and SKM. This algorithm has effective performance in detecting anomalies with various patterns simultaneously.

In CBiForest, SKM is employed initially to group points and then iForest is utilized to calculate anomaly score in each cluster. More details can be found in Algorithm.

Algorithm : CBiForest Algorithm

first step
    employ SKM to cluster all points into two groups
    smaller group is labeled as abnormal
anomaly score of SKM
    distance between anomaly and centroid of bigger cluster
second step
    utilize iForest on two groups to calculate anomaly score
    anomaly score
    point with high score is labeled as abnormal
average path length of iTree
    $c(n) = 2H(n\text{-}1)\text{-}(2(n\text{-}1)/n)$
    where n is the dataset size and H(i) is the harmonic
    number
anomaly score of iForest
    $s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$
    where x is an instance and E(h(x)) is the average of h(x)
    from iTrees
final step
    merge the abnormal results of SKM and iForest together

3. Exploratory Data Analysis

We experiment CBiForest algorithm on the whole wire transaction data during 2016 from CEB. The dataset contains 80 million transaction records from 3.2 million users. For all the experiments, they are conducted in a 32GB Linux server on a distributed environment with three nodes.

Before building the hybrid semi-supervised model, we apply exploratory data analysis in order to have a better understanding of raw data and more inspiration for feature engineering.

3.1. Example Raw Data

By ETL from data warehouse, constructed raw data is presented as Table 1.

TABLE 1. Example Raw Data

Name	Explanation
Event Id	Unique key of transaction events
Transaction Time	Transaction timestamp
Transaction Code	Transaction types
Transaction Status Code	Transaction success or failure
User Id	Unique key of transaction users
Error Message	Reason of transaction failure
Source IP	Transaction IP address
Transaction Channel Code	Transaction Channels

3.2. Transaction Number

First, we investigate transaction number distribution and results are shown in Table 2. Since the user group with less than 6 transactions is not sufficient for fraud detection and the user group with more than 500 transactions is related with particular business scenario in bank, we focus on the user group with transactions between 6 and 500 in Section 5. Filtered dataset contains 40 million transaction records from 1.6 million users.

TABLE 2. Transaction Number Distribution

Transaction Number	User Amount	User Percentage
1	454661	14.64%
2	342906	11.03%
3	368651	11.86%
4	281347	9.05%
5	194826	6.27%
6-10	515260	16.57%
11-50	693721	22.32%
51-100	139730	4.49%
101-500	104186	3.35%
501-1000	8651	0.28%
>1000	4762	0.15%

3.3. Transaction Time

We demonstrate transaction time distribution over month, week and 24 hours. By observation, most transactions are concentrated on weekdays and working hours (Table 3 and Figure 1). These findings contribute to

generate time-related features for detecting anomalies in Section 4.

TABLE 3. Transaction Time Distribution in Week

Date	Transaction Amount	Percentage
Sunday	7152641	8.60%
Monday	14200045	17.07%
Tuesday	13915482	16.72%
Wednesday	13768531	16.55%
Thursday	13001599	15.62%
Friday	13212153	15.88%
Saturday	7960161	9.57%

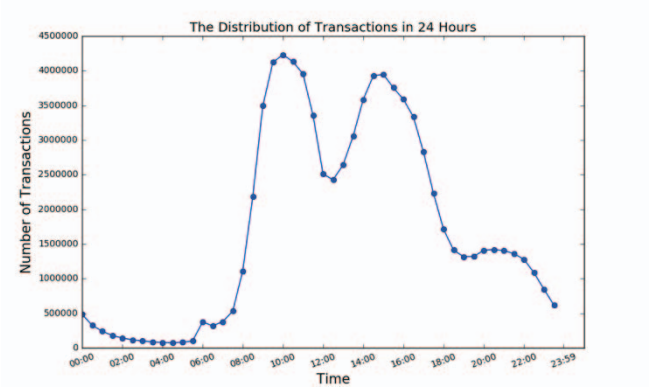

FIGURE 1. Transaction Time Distribution in 24 hours

3.4. Transaction Type

According to our statistics, there are totally 308 wire transaction types in CEB. Based on frequency, we divide transaction types into high type, medium type and low type. Moreover, transaction types are partitioned into transfer, payment and account operation based on transaction purpose. Besides, we summarize risk type and sensitive type in which risk type transactions usually bring more fraud risk in bank and sensitive type is a clustering of transaction types with high failure rate in CEB. All the type divisions are used in Section 4.

3.5. Transaction IP Address

Finally, we research on transaction IP address distribution (Table 4) and find that some users have extremely lots of different IP addresses in a year. This is an obvious anomaly signal and leads to IP-related features in Section 4.

TABLE 4. Transaction IP Address Distribution

IP Number	User Amount	User Percentage
1	1829147	58.84%
2	348056	11.20%
3	172698	5.56%
4	112190	3.61%
5	81982	2.64%
6-10	229501	7.38%
11-50	297149	9.56%
51-100	29369	0.94%
101-500	8575	0.28%
501-1000	20	0.00%
>1000	14	0.00%

4. Feature Engineering

Inspired by the results of exploratory data analysis and previous experience in bank fraud detection, we perform feature engineering in this section. We calculate model metrics in the following 7 aspects: transaction number distribution, time distribution, type distribution, channel distribution, address distribution, periodicity, volatility and fraud pattern. After the work of feature engineering, more than 1000 metrics have been computed.

For the purpose of reducing noise and improving distinction of model, we conduct some methods of feature selection. Correlation matrix and dimensionality reduction are mainly used while 23 important metrics are kept in Table 5 ultimately.

TABLE 5. Selected Model Features

Feature Category	Explanation	Feature Number
Transaction Time Percent	Transaction distribution in week and 24 hours	6
Transaction Type Percent	Distribution of various transaction type divisions	5
Short-term Transaction Number	Include different types of transactions	3
Long-term Transaction Number	Include different types of transactions	3
Midnight Transaction Percent	Include different types of transactions	2
Transaction Number	Include different types of transactions	3
Fraud Pattern	Some Proven Bank Fraud	1

5. Model Building

In this section, we experiment CBiForest algorithm which outperforms individual iForest or SKM in anomaly detection. The methodology is implemented based on filtered 1.6 million users in Section 3B and selected 23 features in Section 4. The whole procedure consists of feature grouping, separate model training, ensemble learning and result analysis.

5.1. Feature Grouping

At the beginning, we plot feature distributions and find an interesting phenomenon that all the features only have two types of distribution shape which are long-tail type and U type (example as Figure 2a and 2b).

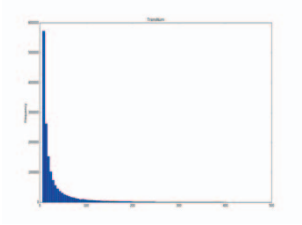

FIGURE 2a. Long-tail Type

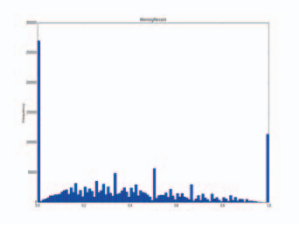

FIGURE 2b. U Type

Considering the characteristics of two algorithms, iForest has better distinction on long-tail type features whereas SKM does on U type features. Thus, we split 23 features into two groups based on their distribution. To maximize the performance of our model, each group of features is trained on its most appropriate algorithm and then combine the anomalies detected together as final result. In this case, CBiForest can have the best distinction on all the features.

5.2. Separate Model Training

For iForest, sub-sampling size and number of trees require to be determined by parameter adjustment. Here sub-sampling size is tested on $[0.1,0.2,0.3,0.4,0.5,0.6,0.7]$ and number of tress is tested on $[100,200]$. We observe that anomaly score curve does not change much no matter what combination of parameters (Figure 3). In this case, we choose sub-sampling size as 0.1 and number of trees as 100.

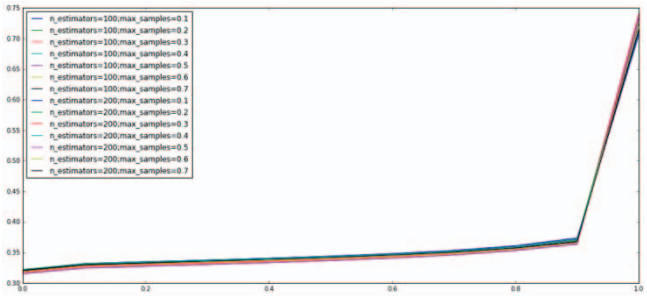

FIGURE 3. Anomaly Score Curve with Parameter Adjustment

For SVM, adjusting parameter is not required. However, we need to create a measurement similar to anomaly score in iForest. Regarding the smaller cluster as anomalies, here we define the anomaly score as the distance between each anomaly and centroid of bigger cluster. In the other word, the farther a sample is away from most users, the more abnormal it is.

According to industry experience, 16000 users (top 1%) are defined as fraudulent in both two algorithms. Furthermore, we compare the distinction of trained iForest between selective 8 long-tail features and all 23 features by measuring the distance of two cluster of users on each feature dimension (example as 4a). We find that selective features have better distinction on trained iForest which proves our idea of feature grouping is the same as trained SKM (example as 4b).

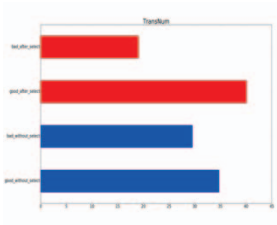

FIGURE 4a. iForest Comparison

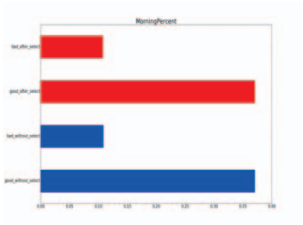

FIGURE 4b. SKM Comparison

(Red represents the result of selective features and blue represents the result of all features)

5.3. Ensemble Learning

Three steps are taken in this learning part. First, we apply trained SKM model to group the whole users into two clusters where the abnormal cluster contains 16000 users and the rest of users is in the normal cluster. Second, trained iForest is performed on both clusters and 16000 users are ranked by anomaly score. Until now, we achieve 16000 abnormal users separately from first two steps where 208 users are labeled by two algorithms simultaneously. Lastly, we merge top 8000 users with highest anomaly score from each step and those 208 users which are 16003 abnormal users totally. More result analysis is shown in the next part.

5.4. Result Analysis

To prove our assumption that CBiForest has effective performance in anomaly detection and outperforms any single algorithm like iForest or SKM, we extract transaction records of 16003 abnormal users and analyze manually by bank worker.

For the results of individual iForest, abnormal users usually have huge volatility during their transaction period or continuous failure transactions over short time. Refer to the example in Table 6, this user tried to purchase financial products several times in two minutes but all failed which

existed the risk of online bank account theft.

TABLE 6. Part of Transaction Records from an Abnormal User of iForest

Transaction Time	Transaction Type	Error Message
20160229 09:50:51	Login	
20160229 09:52:08	Small Foreign Exchange purchasing	
20160229 09:52:27	Small Foreign Exchange Purchasing	
20160301 09:22:53	Login	
20160301 09:25:16	Financial Product Purchasing	Not Supported
20160301 09:26:03	Financial Product Purchasing	Not Supported
20160301 09:27:00	Financial Product Purchasing	Not Supported
20160301 09:27:34	Financial Product Purchasing	Not Supported
20160303 10:50:39	Login	

For the result of individual SKM, abnormal users usually have uneven distribution in their transactions such as extreme amount of payments or midnight operations. For example in Table 7, this user made too many mobile payments with wrong phone number. Clearly, these operations were different from normal transactions.

TABLE 7. Part of Transaction Records from an Abnormal User of SKM

Transaction Time	Transaction Type	Error Message
20160625 07:35:40	Mobile Payment	Phone Number not Matched
20160625 07:35:42	Mobile Payment	Phone Number not Matched
20160625 07:38:56	Mobile Payment	Phone Number not Matched
20160625 07:39:42	Mobile Payment	Phone Number not Matched
20160625 07:43:29	Mobile Payment	Phone Number not Matched
20160625 07:45:38	Mobile Payment	Phone Number not Matched

For the results of CBiForest, not only do they cover the top abnormal users from iForest or SKM but they also find the users with problem in both unstable and unevenly-distributed transaction records. An instance is the user attempting to redeem gift points with others' bank card information (Table 8) which have been verified as fraud in expert system of CEB.

TABLE 8. Part of Transaction Records from an Abnormal User Selected by CBiForest

Transaction Time	Transaction Type	Error Message
20161022 15:43:39	Gift Points Query	
20161022 15:43:49	Gift Points Redeeming	Wrong Expiration Date
20161022 15:44:05	Gift Points Query	
20161022 15:44:15	Gift Points Redeeming	Wrong Expiration Date
20161022 15:44:46	Gift Points Query	
20161022 15:44:56	Gift Points Redeeming	Wrong Expiration Date
20161022 15:45:06	Gift Points Query	
20161022 15:45:18	Gift Points Redeeming	Wrong Expiration Date

Through the above result analysis, we can conclude that our innovative hybrid semi-supervised algorithm does locate abnormal users. The detected users cover various fraud patterns which outperforms the users with fixed fraud characteristic from any individual algorithm.

6. Discussion

Currently, transaction amount and personal information have not been taken into model because of bank security requirement. Besides, we did not make the best of transaction IP address since it would cost over-limit time in mapping IP onto latitude and longitude. These optimizations will be added to our upcoming work for an improved bank fraud detection model.

7. Conclusion

In this paper, we propose a novel hybrid semi-supervised methodology for bank fraud detection in wire transaction which has proved effective in finding abnormal users with various fraud pattern by experiment. Compared with traditional expert system, this machine learning system can reduce false positive rate sharply and learn bank fraud pattern automatically. Furthermore, this model can be utilized in real-time fraud monitoring system and as a supplement for expert system in bank.

Failure to follow the above guidelines may result in a submission being rejected for publication in the conference proceedings and CD ROM.

Acknowledgements

This paper is supported by China Everbright Bank, Beijing Institute of Big Data Research and the IEEE Systems, Man and Cybernetics Society.

References

[1] Frank, M. 2017, "Top 10 Fraud Types for 2017 Based on Losses. Frank on Fraud".

[2] Kokkinaki, A. 1997, "On a typical Database Transactions: Identification of Probable Frauds using Machine Learning for User Profiling", Proc. of IEEE Knowledge and Data Engineering Exchange Workshop; pp 107-113.

[3] Bentley, P., Kim, J., Jung. G. & J Choi. 2000, "Fuzzy Darwinian Detection of Credit Card Fraud", Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society.

[4] Bolton, R. & Hand, D. 2002, "Statistical Fraud Detection: A Review". Statistical Science, 17; pp 235-249.

[5] Dorronsoro, J. Ginel, F. Sanchez, C. & C Cruz. 1997, "Neural Fraud Detection in Credit Card Operations".

IEEE Transactions on Neural Networks, 8; pp 827-834.

[6] Maes, S., Tuyls, K., Vanschoenwinkel, B. & B Manderick. 2002, "Credit Card Fraud Detection using Bayesian and Neural Networks", Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies.

[7] Liu, F. T., Ting, K. M., and Zhou, Z.-H. 2008a, "Isolation Forest", in ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, pp 413–422.

[8] Liu, F. T., Ting, K. M., and Zhou, Z.-H. 2012, "Isolation-based Anomaly Detection", ACM Transactions on Knowledge Discovery from Data, 6 (1); pp 1-39.

[9] Maria, L., ChloéO, A., Guillaume, B., Loïc, L., Aristide, Piwele. 2016, "Credit Card Fraud Detection with Unsupervised Algorithms", Journal of Advances in Information Techonology Vol. 7, No. 1.