Association for Information Systems
AIS Electronic Library (AISeL)

ICEB 2018 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-6-2018

K-Means Algorithm for Recognizing Fraud Users on a Bitcoin Exchange Platform

Yanfeng Wang
School of Business Administration, South China University of Technology, China, bmyfwang@mail.scut.edu.cn

Feng Li
School of Business Administration, South China University of Technology, China, fenglee@scut.edu.cn

Jinya Hu
School of Business Administration, South China University of Technology, China, bmhujinya@mail.scut.edu.cn

Dong Zhuang
School of Business Administration, South China University of Technology, China, dzhuang@scut.edu.cn

Wang, Y.F., Li, F., Hu, J.Y. & Zhuang, D. (2018). K-means algorithm for recognizing fraud users on a bitcoin exchange platform. In *Proceedings of The 18th International Conference on Electronic Business* (pp. 411-420). ICEB, Guilin, China, December 2-6.

K-Means Algorithm for Recognizing Fraud Users on a Bitcoin Exchange Platform

(Full paper)

Yanfeng Wang*, School of Business Administration, South China University of Technology, China, bmyfwang@mail.scut.edu.cn
Feng Li, School of Business Administration, South China University of Technology, China, fenglee@scut.edu.cn
Jinya Hu, School of Business Administration, South China University of Technology, China, bmhujinya@mail.scut.edu.cn
Dong Zhuang, School of Business Administration, South China University of Technology, China, dzhuang@scut.edu.cn

ABSTRACT

This paper addresses recognizing fraud users on a Bitcoin exchange website-bitcoin-otc. According to online rating records provided by the website, some users behave significantly different from others. Seeing that, the classical K-means clustering algorithm is proposed to identify these abnormal users. K-means algorithm is an unsupervised clustering algorithm that clusters users based on feature similarity. Therefore, performance of K-means algorithm relies on the features. This paper explored and found the best collection of features based on real record data, e.g., mean of total ratings sent. Since the selected features are not observed for record set, the website should offer these features for potential traders.

Keywords: K-means algorithm, P2P website, Fraud detection.

*Corresponding author

INTRODUCTION

On 1 January 2018, the price of Bitcoin was 13,412USD (as shown in Figure 1), which is 13 times compared with the price of the same day of 2017 (997USD). It increasingly attracts global investors to Bitcoin marketplace. However, Bitcoin is a decentralized digital currency that is transacted directly between users without any intermediation. As such, trading platforms mostly provide no authentication service, and the users are exposed to counterparty risk.

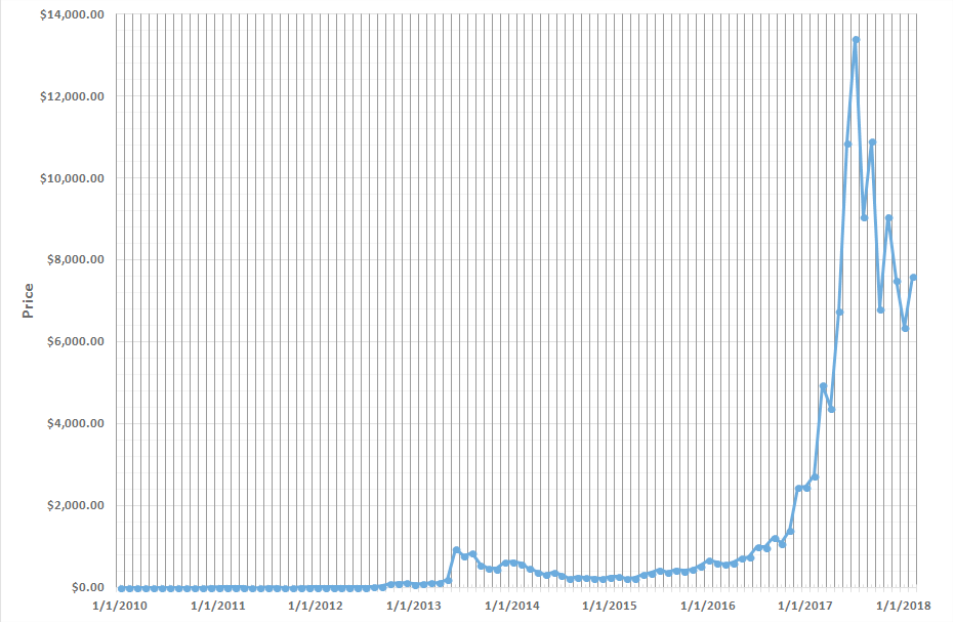

Figure 1: Line Chart of Bitcoin Price (from U.S. Finance Reference, https://www.officialdata.org/bitcoin-price)

To help users reduce trading risk, the websites usually offer a rating system along with the trading platform. On this system, users can give others ratings after transactions. For example, positive ratings mean that user trusts this person, while negative ratings for a person as a fraudster. It is suggested that user access to his counterparty's trade and reputation history firstly to avoid dealing with fraudster. The question is that the ratings are not foolproof. Fraudulent users can also inter-rate each other with positive ratings. As a result, normal users need to stay vigilant to be prey for fraudster.

This paper is motivated by the observation that behaviors of fraudulent users seem to be quite different from normal users. For instance, a normal user often received '+1' ratings from the counterparty, while a fraudster received '+10' ratings. So, we apply a classical clustering algorithm- K-means algorithm to recognize abnormal users from others. We further study the features for the K-means algorithm. A collection of features are chosen with the best performance of K-means algorithm.

This paper is organized as follows. Section 2 reviews the literature and K-means algorithm is described in Section 3. Then, the selection of features for the K-means algorithm is presented based on real data. Section 5 shows the results of data analysis. The last section concludes the whole paper.

RELATED WORK

When the Bitcoin was firstly invented as a form of electronic cash in November 2008, the price of bitcoin was nearly zero. Even on 1 January 2011, the price is 0.30USD per bitcoin. At that time, Bitcoin was argued to be not a currency rather a speculative asset (Corbet, Lucey, Peat, & Vigne, 2018). It was also beyond all people's imagination that the price is rising to 13412.44USD in 2018.

Since the price of Bitcoin rises exponentially, it has been a frequent target of attacks by financially-motivated criminals (Gandal, Hamrick, Moore, & Oberman, 2018). Researchers, therefore, focused on security and fraud detection of the Bitcoin ecosystem. For example, Ziegeldorf, Matzutt, Henze, Grossmann, and Wehrle (2018) proposed a novel oblivious shuffle protocol improves resilience against malicious attackers. Kumar, Spezzano, Subrahmanian, and Faloutsos (2016) represented Bitcoin transaction as a weighted signed network. In the network, edges are labeled with positive or negative weights to present the ratings the rater sent the ratee. They use the idea of HIT algorithm to measure how much a node is trusted by other nodes (Kleinberg, 1999). Maesa, Marino, and Ricci (2017) treated Bitcoin transaction as a simply graph and discovered that the graph was not a small world network with some unusual patterns. They explained that these patterns are probably due to artificial users behaviors. Kumar et al. (2018) proposed a REV2 algorithm to improve the prediction of fraudulent users. In the paper, fraud detection is categorized into network-based fraud detection algorithm (Akoglu, Tong, & Koutra, 2015) and behavior-based fraud detection algorithm (Jiang, Cui, & Faloutsos, 2016).

The applied algorithm in this paper belongs to behavior-based fraud detection algorithms, namely the K-means algorithm. K-means algorithm, sometimes K-means clustering called, is a very popular and efficient algorithm for clustering analysis (Zhao, Deng, & Ngo, 2018). For example, the K-means algorithm was introduced to the recommender system (Kant et al., 2018), fashion design (Vincent, Makinde, Salako, & Oluwafemi, 2018), social network analysis (Liu, Ma, Xiang, Tang, & Zhang, 2018), network evolution (Yang & Chen, 2018) etc. Therefore, the K-means algorithm is adopted here to detect fraudsters. To improve the performance of K-means algorithm, initial cluster centers assignment is very important (Karegowda, Vidya, Jayaram, & Manjunath, 2013) (Sirait & Arymurthy, 2011). But, feature selection is another underestimated key to the success of K-means clustering (Mavroeidis & Marchiori, 2014). For example, linear discriminant analysis (LDA) with trace ratio criterion is used in the research (Wang, Nie, & Huang, 2014). This paper also addresses the feature selection problem. Based on the idea from Kumar et al. (2018), we explored a bundle of features to match the real data best, including features from network-based fraud detection algorithm, and behavior-based fraud detection algorithm.

BITCOIN TRADING PLATFORMS AND K-MEANS ALGORITHM

Bitcoin Trading Platform- Bitcoin OTC

Bitcoin OTC (https://www.bitcoin-otc.com) is an online marketplace for bitcoin trading. All transactions occurred on this platform are conducted directly between counterparties, without any guarantee from platform. All exchange risks are taken by buyers and sellers. To help users decrease trading risk (one of the parties did not pay), the platform offers a service named 'web of trust,' where users can access counterparty's reputation and trade history.

In this web database, the basic record of the user is:

id	rater nick	rater total rating	rated nick	created at (UTC)	rating	notes
		1061		2010-11-30 23:31:51	1	
		614		2011-02-21 05:49:06	7	
		481		2013-12-16 21:26:35	1	nanooooo!
		439		2014-04-24 03:57:20	5	.
		377		2014-02-23 05:34:06	4	several deals..Smooth™

(a) List of All Ratings the Ratee Received

id	nick	first rated (UTC)	keyid	total rating	number of positive ratings received	number of negative ratings received	number of positive ratings sent	number of negative ratings sent
		2010-11-08 18:14:09		801	226	0	206	9
		2010-11-08 18:35:30		95	39	3	43	2
		2010-11-08 18:35:39		-16	12	10	0	0
		2010-11-08 18:36:00		168	54	0	60	3

(b) List of the Aggregate Ratings of Ratee

Figure 2: A Snapshot of 'Web of Trust' on the Bitcoin-otc Website

The ratings are between -10 and +10 (integral value, and could not be set to 0). The higher score means that the ratee has better reputation in the exchange, and the lower score is the opposite. For example, according to a guideline of the system, a ratee is marked as '5' if 'you've had some good transactions with this person,' '-10' if 'person failed to hold up his end of the bargain, took payment and ran, fraudster.' Therefore, if a user is rated positively by most people he had traded with, he could be labeled as a trust-worthy user. Once a user was rated negatively by other people, the most time he would often be labeled as a fraudster. Generally, users on bitcoin-otc platform follow these intuitions and opinions:

Table 1: The Intuitions And Opinions of Evaluation

Ratings	**Intuitions and opinions**
Positive 'rating' value and high 'total rating' value	A trusted user, low trading risk
Low 'total rating' value and existing negative 'rating' value	Fraudsters, trading risk
Existing both positive and negative 'rating' value	Uncertain, high trading risk

However, these intuitions are not always right. Some fraudsters created several accounts who rated each other with high rating values. When other users browser the 'web of trust' database, 'total rating' value of the user is a large number. In this way, an innocent user often falls prey to fraudsters groups.

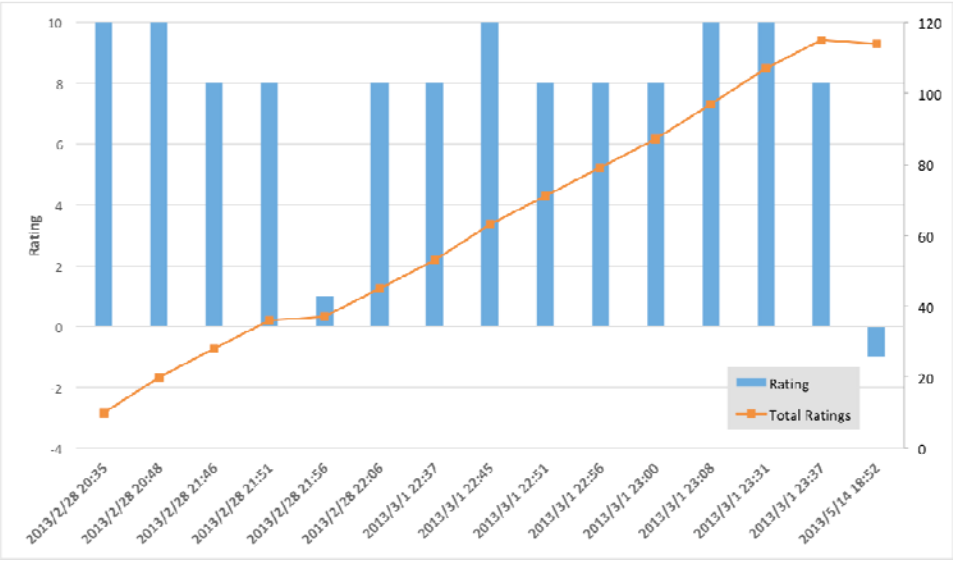

Figure 3: A Line Chart of Rating Time Series for a Typical 'Scammer'

As shown in figure 3, the user was rated '10' or '8' by 13 users in two days (from 28 February to 1 March). After that, the user got very high scores-115. Then, he was reported as a fraudster.

Comparing with the behavior of the fraudster in figure 3, a normal user's behavior is quite different (as shown in figure 4).

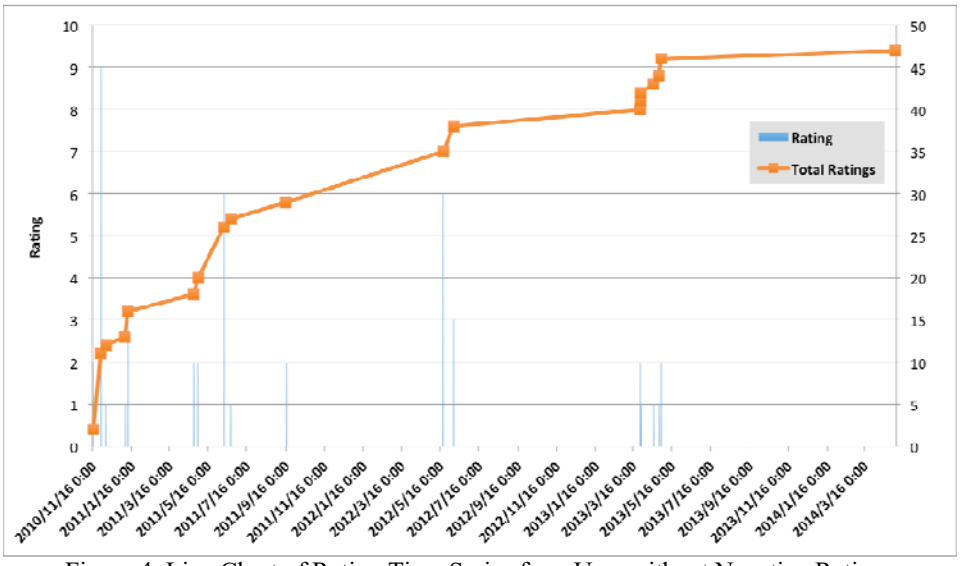

Figure 4: Line Chart of Rating Time Series for a User without Negative Ratings

In figure 4, the user got only 47 ratings in three and a half years. And the average rating received was 2.47.

Social Network Analysis Of The Dataset

This paper takes an open dataset called 'Bitcoin-OTC' collected by the network analysis project of Stanford University (http://snap.stanford.edu). The dataset contains 35,592 rating records among 5,881 users.

Firstly, a simple graph is built from the dataset, where each node represents a user, and each edge represents existing rating from a rater to a ratee.

With the social network analysis tool Pajek 5.05 (http://mrvar.fdv.uni-lj.si/pajek/), the basic statistical information is shown in the following table 2 (Kiss & Bichler, 2008):

Table 2: The Basic Information of the Rating Network

Feature	Definition	Evaluation
Total number of nodes	Total number of users	5,881
Total number of edges	Total number of rating of users	35,592
Average in-degree	The average number of ratings received by a node	6.0520
Average out-degree	The average number of ratings sent by a node	6.0520
Average clustering coefficient	The degree to which nodes tend to cluster together	0.2416
Network diameter	The longest distance between any two nodes	11
Average path length	The average of all the minimum path lengths between all pairs of nodes in a network	3.7189

K-means Clustering Algorithm

K-means algorithm is a kind of clustering analysis algorithm, which divides *n* sample points into *k* clusters. Sample points within cluster have high similarity, while sample points between clusters have low similarity. Similarity calculation is based on the average value of sample points in a cluster. The process of the algorithm is as follows:

1）In the dataset *D*, randomly select *k* points and assigned them to the initial cluster center, $C_i = P_k, (i = 1, \cdots, k)$;

2）For all $P_j, (j = 1, \cdots, n)$ in *D*, calculate its distance to each cluster center $C_i, (i = 1, \cdots, k)$;

$$d(P_j, C_i) = \sqrt{\sum_{i=1}^{m} (P_j^i - C_i^i)^2} \quad (1)$$

In (1), $P_j^l, C_i^l, (l-1,\cdots,m)$ are the features of sample points.

3）Find out the minimum distance between P_j and C_i, $C_a = \min\ d(P_j, C_i), (i = 1,\cdots,k)$, and reassign P_j to the cluster C_a;

4）Recalculated the features of clustering center $C_i^l, (l = 1,\cdots,m; i = 1,\cdots,k)$ based on the total J sample points in the cluster:

$$C_i^l = \sum_{j=1}^{J} P_j^l / J, (i = 1,\cdots,k; l = 1,\cdots,m) \quad (2)$$

5）Repeat step 2 - 4, until features of all center of clusters do not change anymore.

SELECTION OF FEATURES FOR K-MEANS ALGORITHM

Pretreatment Of Users

According to the analysis of users' behavior in Table 1, if a user had been rated with positive values by all users, it implied that the user was trusted. On the contrary, if a user had ever been rated with negative values by other users, it implied that the user was a fraudster. But due to existing of some scammers rated by other fraudsters with high ratings (as shown in Figure 3), not all users rated with positive values are really trusted. Seeing that, we use the K-means algorithm to separate 'suspects' from 'trusted' users. Before the K-means algorithm is used to classify suspect users from trusted users, we filtered those users that were confirmed as fraudsters (rated with negative values).

Table 3: The Pretreatment of Nodes

Name	Character	Number
Positive-class	nodes that have been rated with positive values by other users	4,604
Negative-class	nodes that have been rated with negative values by other users	1,254
Zero-class	nodes that have not been rated by other users	23

As shown in Table 3, 1277 nodes were removed from dataset because they were labeled as fraudsters by others.

Selection Of Features

Through the study of website users, this paper finds that it is not always effective to judge the credibility of accounts simply based on the score that they got. As shown in figure 3, for the designated user, between February 28 and March 1, 2013, 14 users mainly gave it a high score like +8 and +10. Nearly two months later, the user was scored '-1' for its abnormal behavior. However, its mean score given by other users is 8.2143 (115/14), which looks like an honest user. To more effectively determine whether users have exchange risks, this paper identifies the following 12 indexes analyzing the Positive-class nodes according to the three fields (source, target, rating) in the dataset.

Table 4: The Definition of Index

Number	Index	Definition
1	A1	number of total ratings sent
2	A2	number of positive ratings sent
3	A3	number of negative ratings sent
4	B1	number of total ratings received
5	B2	number of positive ratings received
6	B3	number of negative ratings received
7	C1	mean of total ratings sent
8	C2	mean of positive ratings sent
9	C3	mean of negative ratings sent
10	D1	mean of total ratings received
11	D2	mean of positive ratings received
12	D3	mean of negative ratings received

Classification And Characteristic Description Of Positive-class Nodes

To further discover the overall characteristics of Positive-class nodes, this paper excavates the characteristics of nodes which are significantly different from other user nodes.

It can be seen from the above analysis that the Positive-class nodes contain abnormal nodes, and their risks are mainly reflected in the abnormal score. Therefore, in this paper, the Positive-class nodes with abnormal 'mean of total ratings sent/received' will be regarded as abnormal nodes. Abnormal nodes are likely to be nodes with exchange risk.

According to the statistical information, the average of C1-index of Positive-class nodes is 1.5393, and the standard deviation is 1.8965. Therefore, Positive-class nodes are divided adopting triple standard deviations (7.2289=1.5393+3×1.8965) and the nodes whose value of C1-index is greater than 7.2289 are screened out as the abnormal users. For the convenience of description, such nodes are defined as Positive1-class nodes. Similarly, Positive-class nodes are divided adopting triple standard deviations (=5.5038) according to the average (=1.6653) and standard variance (=1.2974) of D1-index. For the convenience of description, such nodes are defined as Positive2-class nodes. The basic statistical information of these four types of nodes is shown in table 5.

Table 5: The Statistical Information of Positive-class Nodes

		Positive0	Positive1	Positive2	Positive12
Number of nodes		4,425	113	104	38
A1	Average	4.4573	1.2124	1.4327	1.2368
	Standard deviation	16.9492	0.7252	1.6474	0.6339
A2	Average	4.2328	1.2124	1.3462	1.2368
	Standard deviation	15.6425	0.7252	1.5122	0.6339
A3	Average	0.2245	0	0.0865	0
	Standard deviation	3.6398	0	0.6983	0
B1	Average	4.3590	1.2920	1.5865	1.3158
	Standard deviation	13.4024	0.7755	1.0391	0.6619
B2	Average	4.3590	1.2920	1.5865	1.3158
	Standard deviation	13.4024	0.7755	1.0391	0.6619
B3	Average	0	0	0	0
	Standard deviation	0	0	0	0
C1	Average	1.3168	9.5246	5.2284	9.5394
	Standard deviation	1.3674	0.8317	4.2028	0.8251
C2	Average	1.4665	9.5246	5.3820	9.5394
	Standard deviation	1.1901	0.8317	3.9266	0.8251
C3	Average	-0.5590	0	-0.2019	0
	Standard deviation	2.1635	0	1.3821	0
D1	Average	1.5198	4.0893	7.7972	8.8026
	Standard deviation	0.8342	3.6123	1.8900	1.5314
D2	Average	1.5198	4.0893	7.7972	8.8026
	Standard deviation	0.8342	3.6123	1.8900	1.5314
D3	Average	0	0	0	0
	Standard deviation	0	0	0	0

After analysis, the following conclusions can be drawn from table 5:

(1) The value of A1-index, A2-index, B1-index, and B2-index of Positive1-class nodes and Positive2-class nodes are low. Their average of these four indexes are close to 1-2, but those of Positive0-class nodes are between 4 and 5.
(2) The value of C1-index and C2-index of Positive1-class nodes and Positive2-class nodes are relatively high. Their average of these two indexes are both greater than 5, but those of Positive0-class nodes are between 1 and 2.
(3) The value of D1-index and D2-index of Positive1-class nodes and Positive2-class nodes are relatively high. Their average of these two indexes are both greater than 4, but those of Positive0-class nodes are between 1 and 2.

This indicates that the behaviors of Positive1-class nodes and Positive2-class nodes are significantly different from other nodes. The frequency of total ratings sent/received of these two types of nodes is low, but the mean of total ratings sent/received of them is high.

RECOGNITION OF POSITIVE1-CLASS NODES

Research Approach

This paper applies the k-means algorithm to classify the Positive-class nodes and finally finds out the indexes which can identify Positive1-class nodes accurately.

First, input the original data and select different index combinations according to the actual situation to divide the Positive-class nodes into four categories by using the k-means algorithm. Then compare the original Positive1-class nodes with the classification

results to find the combinations of indexes with higher recall and precision. Finally, compare the indexes to obtain identification enlightenment of Positive1-class nodes.

Data Analysis

14 combinations of indicators with high recognition precision are listed below, and they are compared in groups.

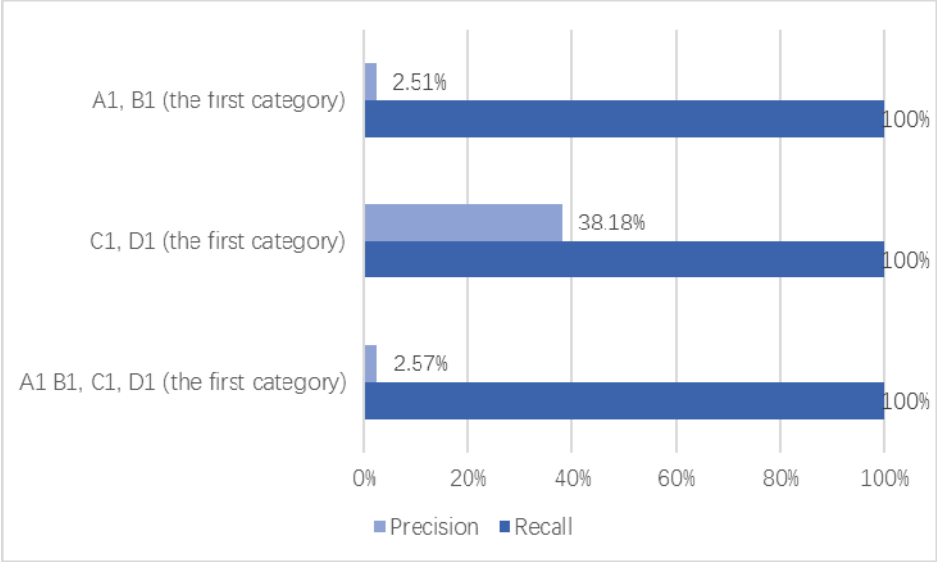

Figure 5: Comparison of the Recognition Effect Between 'A1, B1' and 'C1, D1'.

As shown in figure 5, on the premise that the recall is both 100%, the precision of the combination of A1-index and B1-index is only 2.51%, while the precision of the combination of C1-index and D1-index is 38.18%. Besides, the addition of A1-index and B1-index will reduce the precision of recognition of Positive1-class nodes. The precision of the combination of C1-index and D1-index is 38.18%, and after the addition of A1-index and B1-index, the precision is only 2.57%.

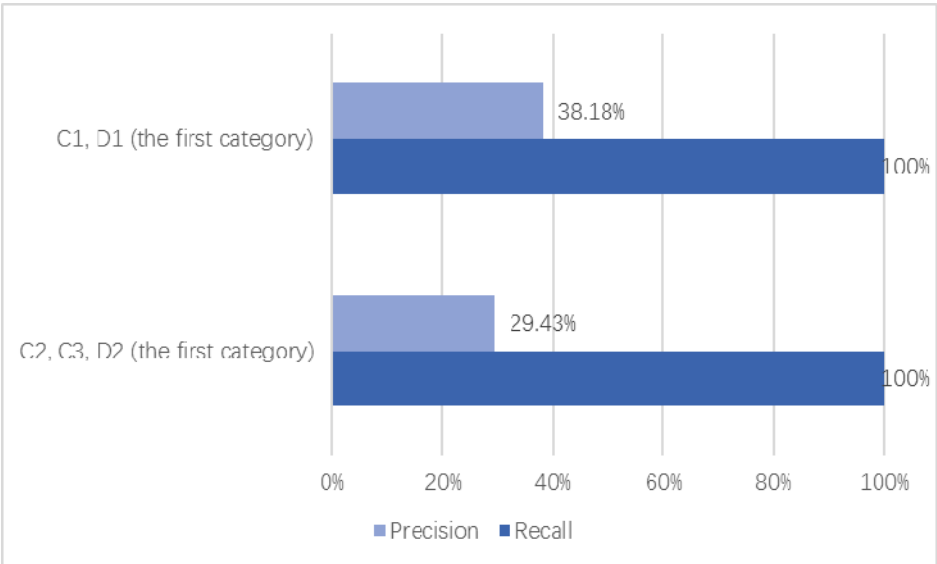

Figure 6: Comparison of the Recognition Effect Between 'C1, D1' and 'C2, C3, D2'.

As shown in figure 6, on the premise of the recall with 100%, the precision of the combination of C1-index and C2-index is 38.18%, while the precision of the combination of C2-index, C3-index and D2-index is 29.43%. This indicates that the precision of the combination of total mean indexes (C1, D1) is higher than that of the combination of positive (C2, C3) and negative mean indexes (D2).

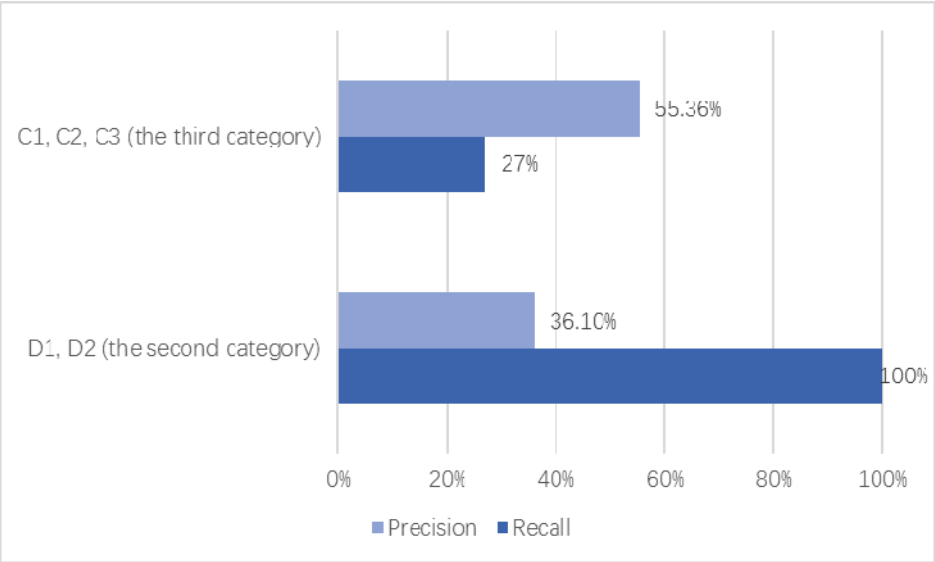

Figure 7: The Comparison of Recognition Effect Between 'C1, C2, C3' and 'D1, D2'.

As shown in figure 7, under the condition that the recall is both 100%, the precision of the combination of C1-index, C2-index and C3-index is 36.1%. The recall of the combination of D1-index and D2-index is low, only 27%, but its precision is 55.36%. As the precision of identification is more important in the practical application, the precision is mainly considered here. It can be seen that the recognition precision of 'ratings sent indexes' is lower than 'ratings received indexes'.

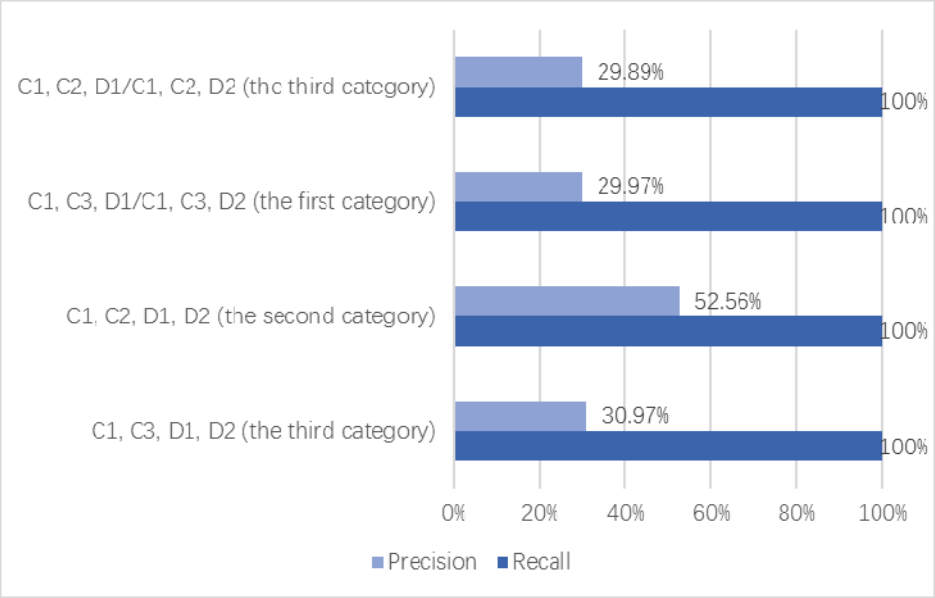

Figure 8: The Comparison of Recognition Effect Between 'C2' and 'C3'.

As shown in figure 8, on the premise that the recall rate is both 100%, the precision of the combination of C1-index, C2-index and D1-index/C1-index, C2-index and D2-index is 29.89%, and that of the combination of C1-index, C3-index, and D1-index/C1-index, C3-index and D2-index is 28.97%. The precision of the combination of C1-index, C2-index, D1-index, and D2-index is 52.56%, and that of the combination of C1-index, C3-index, D1-index, and D2-index is 30.79%. It can be seen that the C2 index is more accurate than C3 in recognition. That is to say, comparing to the negative score sending to others, the positive score sending to others has higher precision in identification.

Finally, we combine indicators related to the mean score (C1, C2, C3, D1, D2) to identify Positive1-class nodes and find that this combination of indexes has the highest recall and precision.

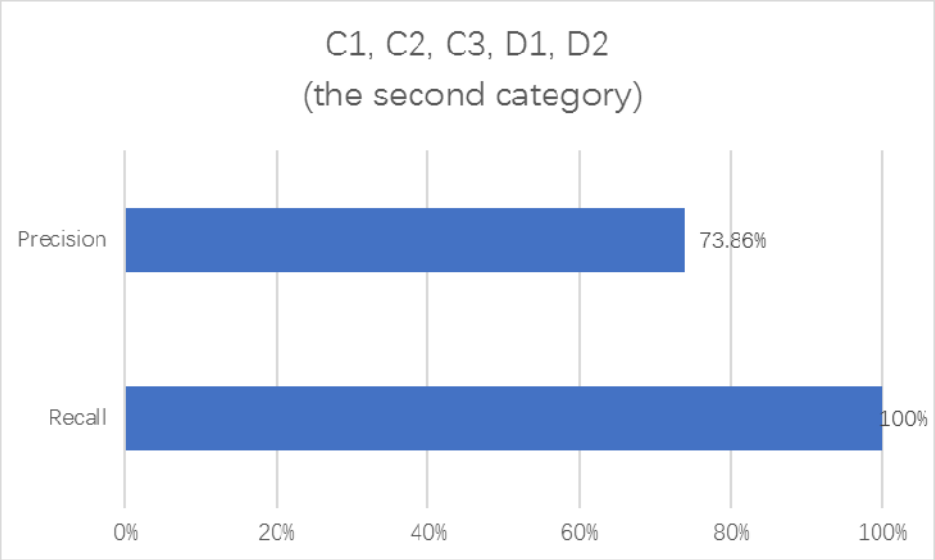

Figure 9: The Results of the Recognition of 'C1, C2, C3, D1, D2'.

Conclusion Of Identification

Based on the above data analysis, the following conclusions can be drawn:

Conclusion 1: In the identification of Positive1-class nodes, the number of total ratings sent/received is not the accurate recognition indexes.

Conclusion 2: In the identification of Positive1-class nodes, compared with the 'mean of negative ratings sent' indexes, the 'mean of positive ratings sent' indexes have more reference value.

Conclusion 3: In the identification of Positive1-class nodes, the precision of the combination of total mean indexes is higher than that of the combination of only positive or negative mean indexes. If ignore the positive and negative mean of ratings, only using the total mean of ratings to identify will reduce the recognition accuracy.

Conclusion 4: In the identification of Positive1-class nodes, the combination of mean indexes has the highest recognition accuracy.

CONCLUSIONS

Through the positive and negative scores between users, this paper finds the problem nodes which are different from the normal nodes. Then, the behavior characteristics of these abnormal nodes are found. The most important thing is that this paper finds the mean index combination (C1, C2, C3, D1, D2) which can recognize nodes having potentially risk accurately. It also certifies that the K-means algorithm is effective in recognizing 'potential fraud' nodes.

ACKNOWLEDGMENT

This work is partially supported by the Education and Teaching Reform Project of the South China University of Technology (Y9161010).

REFERENCES

[1] Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, *29*(3), 626-688.

[2] Corbet, S., Lucey, B., Peat, M., & Vigne, S. (2018). Bitcoin Futures—What use are they?. *Economics Letters*, *172*, 23-27.

[3] Gandal, N., Hamrick, J. T., Moore, T., & Oberman, T. (2018). Price manipulation in the Bitcoin ecosystem. *Journal of Monetary Economics*, *95*, 86-96.

[4] Jiang, M., Cui, P., & Faloutsos, C. (2016). Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems*, *31*(1), 31-39.

[5] Kant, S., Mahara, T., Jain, V. K., Jain, D. K., & Sangaiah, A. K. (2018). LeaderRank based k-means clustering initialization method for collaborative filtering. *Computers & Electrical Engineering*, *69*, 598-609.

[6] Karegowda, A. G., Vidya, T., Jayaram, M. A., & Manjunath, A. S. (2013). Improving performance of k-means clustering by initializing cluster centers using genetic algorithm and entropy based fuzzy clustering for categorization of diabetic patients. In *Proceedings of International Conference on Advances in Computing* (pp. 899-904). Bangalore, India, July 4-6.
[7] Kiss, C., & Bichler, M. (2008). Identification of influencers—measuring influence in customer networks. *Decision Support Systems*, *46*(1), 233-253.
[8] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604-632.
[9] Ziegeldorf, J. H., Matzutt, R., Henze, M., Grossmann, F., & Wehrle, K. (2018). Secure and anonymous decentralized Bitcoin mixing. *Future Generation Computer Systems*, *80*, 448-466.
[10] Kumar, S., Spezzano, F., Subrahmanian, V. S., & Faloutsos, C. (2016). Edge Weight Prediction in Weighted Signed Networks. In *Proceedings of 2016 IEEE 16th International Conference on Data Mining* (pp. 221-230). Barcelona, Spain, December 12-15.
[11] Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., & Subrahmanian, V. S. (2018). Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 333-341). Marina Del Rey, CA, February 5-9.
[12] Liu, H. L., Ma, C., Xiang, B. B., Tang, M., & Zhang, H. F. (2018). Identifying multiple influential spreaders based on generalized closeness centrality. *Physica A: Statistical Mechanics and its Applications*, *492*, 2237-2248.
[13] Maesa, D. D. F., Marino, A., & Ricci, L. (2017). Detecting artificial behaviours in the Bitcoin users graph. *Online Social Networks and Media*, *3*, 63-74.
[14] Mavroeidis, D., & Marchiori, E. (2014). Feature selection for k-means clustering stability: theoretical analysis and an algorithm. *Data Mining and Knowledge Discovery*, *28*(4), 918-960.
[15] Sirait, P., & Arymurthy, A. M. (2010). Cluster centres determination based on KD tree in K-Means clustering for area change detection. In Proceedings of 2010 *International Conference on Distributed Framework for Multimedia Applications* (pp.1-7). IEEE, Yogyakarta, Indonesia, August 2-3.
[16] Vincent, O. R., Makinde, A. S., Salako, O. S., & Oluwafemi, O. D. (2018). A self-adaptive k-means classifier for business incentive in a fashion design environment. *Applied computing and informatics*, *14*(1), 88-97.
[17] Wang, D., Nie, F., & Huang, H. (2014, September). Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In Proceedings of *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 306-321). Skopje, Macedonia, September 18-22.
[18] Yang, Z., & Chen, X. (2018). Evolution assessment of Shanghai urban rail transit network. *Physica A: Statistical Mechanics and its Applications*, 503, 1263-1274.
[19] Zhao, W. L., Deng, C. H., & Ngo, C. W. (2018). K-means: A revisit. *Neurocomputing*, *291*, 195-206.