2024 International Conference on Intelligent Computing and Data Analytics (ICDA) | 979-8-3315-4194-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDA64850.2024.00025

Financial Data Anomaly Detection Based on Data Mining Technology

Fen Song *
Yantai Vocational College, yantai, 264670, China
* yantaisongfen@163.com

***Abstract*—In order to understand the anomaly detection of financial data, this paper proposes a research method for anomaly transaction detection based on data mining, which can detect anomalies in transactions at both the business and operational levels. Firstly, when a user submits a new consumption transaction, this paper uses Bayesian belief network algorithm to determine the posterior probability that the current transaction belongs to a normal transaction, as the trust factor at the business level; Then extract several operations of the user before the current transaction, and together with the current transaction, form a fixed length operation sequence. Use the BLAST-SSAHA algorithm to compare it with the user's normal operation sequence and known abnormal operation sequence, and obtain the credibility factor at the operation level. Taking into account both the credibility factor at the business level and the credibility factor at the operational level, the final decision is made on whether the current transaction is an abnormal transaction.**

Keywords-Data mining; Financial data; Anomaly detection

I. Introduction

With the continuous development of China's economy, as the core of the financial market system, the number and scale of transactions in banks are gradually increasing, leading to various illegal transactions such as money laundering and illegal fundraising being repeatedly prohibited. These illegal transactions seriously disrupt the national financial order and harm the interests of the people, so it is increasingly important to detect and crack down on illegal trading activities. How to effectively detect abnormal transactions in banks from massive transaction data and timely crack down on illegal and criminal transaction behaviors is a common challenge faced by various banks. A mathematical model needs to be constructed to analyze transaction data, and then use computers to efficiently screen out accurate abnormal transaction information. In abnormal transactions, such as illegal fundraising transactions, the characteristic is to transfer funds from different accounts to another account within a certain period of time, which creates clustering in transaction information. This clustering is non random, and by detecting these non random clustering transaction information, abnormal transaction information suspected of illegal activities can be screened.

Abnormal transaction detection is usually based on two assumptions: firstly, there is a significant difference between abnormal transactions and normal transactions; Another reason is that the proportion of abnormal transactions in all transactions is very small. According to different detection principles, abnormal transaction detection techniques mainly include statistical methods, bias based methods, and density based methods.

Statistical methods first model data points using a certain distribution (such as normal distribution, Poisson distribution, etc.), and then use inconsistency tests to determine anomalies. The limitation of this method is that the data distribution in reality often does not conform to any known ideal distribution; In addition, most tests are focused on individual attributes, and the effectiveness of anomaly detection in multidimensional data is not ideal. The deviation based method identifies abnormal data by checking a set of object features, and objects that deviate from the given description are defined as anomalies. The main technique used in this method is sequence anomaly technique, which imitates human thinking patterns and discovers elements that are different from most data from a set of continuous sequences. The density based method introduces the concept of Local Outlier Factor (LOF), which measures the degree of anomaly of an object with respect to its surrounding neighbors and can detect data with local anomalies[1-2].

II. Abnormal transaction detection method based on data mining

A. Design Architecture

The method we propose for detecting abnormal transactions mainly involves comparing the current transaction with the user's past transaction records. The main idea is to comprehensively judge whether the user's transaction is abnormal based on the user's current submitted consumption transactions and their recent operation records. This method mainly includes two stages: deployment and detection. The deployment phase mainly involves analyzing the user's past transactions and operations, as well as learning the classifier; The detection stage is the process of initiating a classifier and sequence aligner to perform anomaly detection on a consumer transaction after it is detected. Finally, this section will provide the credibility factors of the current transaction at the business and operational levels, and combine these two factors to determine whether there has been an anomaly in the transaction. We choose consumer transactions to trigger anomaly detection function mainly because consumer transactions are the most frequent transactions in which users are involved in fund flow, and are more important compared to other transactions. If anomaly detection is performed on all transactions, it will undoubtedly lead to a decrease in the performance of the entire payment system.

In business level anomaly detection, we use Bayesian Belief Network (BBN) to classify current transactions. For each transaction of a user, we select transaction time, transaction location, transaction amount, and merchant type as feature vectors[3-4].

979-8-3315-4194-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDA64850.2024.00025

In the anomaly detection at the operational level, we use the BLAST-SSAHA algorithm to construct a sequence of N length operations based on the user's current consumption operations and previous N-1 operations. This sequence is then used as the query sequence for the BLAST-SSAHA algorithm, and compared with the user's previous normal operation sequences in the database. We then compare it with known abnormal operation sequences in the database to determine their similarity with these sequences. Figure 1 shows the architecture of our abnormal transaction detection method.

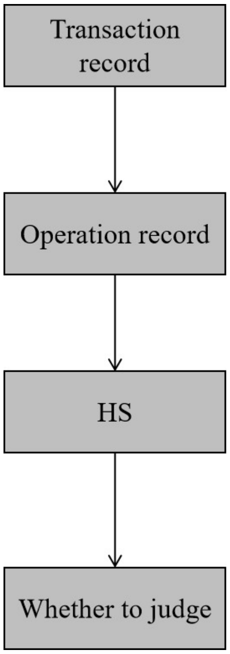

Figure 1. Architecture of Abnormal Transaction Detection Methods

III. Business level anomaly detection

A. *k-means clustering algorithm*

In our proposed method, the k-means algorithm is used to discretize the values of transaction position and transaction amount in each transaction for the use of Bayesian belief network classifiers[5-6]. The K value represents the number of clusters in the k-means algorithm, and selecting the appropriate K value is crucial for the effectiveness of data discretization. If the K value is too small, the clustered data points will appear too general, which may cause the differences between different categories to become blurred and reduce the recognition accuracy of the Bayesian belief network classifier. On the contrary, if the K value is too large, the data points will be divided too finely, increasing the complexity of the calculation, and may also introduce noise, which is not conducive to the classifier's generalization ability.

The k-means algorithm can divide all samples into k categories based on the distance between samples in the dataset, and the value of k is specified by the user in advance. The core idea of this algorithm is to minimize the sum of distances from all points in the classification to their respective classification center points. Let p represent a point in a certain classification, Cn represent the nth classification, mn represent the center point of Cn, the function D (p, mn-n) represents the distance between point p and point m, which is usually calculated using Euclidean distance. The k-means algorithm will find a partitioning method that minimizes the squared error E, as shown in equation (1):

$$\mathrm{E} = \sum_{\mathrm{i}=2}^{\mathrm{K}} \sum_{\mathrm{peci}} D(p, m_{\mathrm{n}}) \qquad (1)$$

At the beginning of the algorithm, m points are randomly selected as the center points for each classification, and then iteration begins. Each iteration consists of the following two steps:

(1) Cluster each non center point, and each point will be assigned to the classification represented by the closest center point.

(2) To recalculate the center points for the k well divided categories, the method of calculating the average is generally adopted, and the point closest to the average is selected as the new center point.

When there is no change in the classification of any non center point in an iteration, or when the predetermined number of times is reached, the iteration process ends. At this point, the dataset is divided into k categories.

B. *Bayesian Belief Network Classifier*

In our proposed method, the BBN algorithm is used to classify the user's current consumption transactions and calculate their posterior probability of belonging to normal transactions. BBN is a machine learning algorithm based on probabilistic graph models, which excels at handling data with uncertainty and causal relationships. When conducting anomaly detection at the business level, transaction data often has multiple influencing factors, such as transaction location, transaction amount, time, etc., and there may be complex dependencies between these factors. BBN can model these complex dependency relationships and make probabilistic inferences based on existing data to identify potential abnormal transactions. Therefore, BBN has been chosen for anomaly detection at the business level, which can better adapt to complex business environments with multiple factors and levels.

BLAST-SSAHA algorithm: BLAST-SSAHA is an efficient sequence alignment algorithm originally used for gene sequence alignment in bioinformatics. When conducting anomaly detection at the operational level, the focus is on quickly processing large amounts of transaction data and identifying abnormal behaviors that are significantly different from known patterns. BLAST-SSAHA can efficiently identify abnormal transaction patterns by quickly matching transaction sequences with historical data. This algorithm has the ability to process large-scale data and achieve fast comparison, making it unique in real-time anomaly detection at the operational level. The selection of BBN and BLAST-SSAHA as the main algorithms for anomaly detection is based on their respective advantages and applicable scenarios. BBN is suitable for complex reasoning and analysis at the business level, and can accurately identify abnormal behaviors under the influence of multiple factors; BLAST-SSAHA is suitable for fast and real-time anomaly detection at the operational level, ensuring the security and stability of the system in high concurrency situations. This combination of double-layer surfaces and dual algorithms can effectively improve the accuracy and efficiency of overall anomaly detection.

The principle of the BBN classification algorithm is to calculate the posterior probability of the current sample belonging to each classification under the condition of a given observation value, and select the category with the highest posterior probability as the classification result of the current sample. This can be seen as an improvement on the naive Bayesian classification algorithm.

A sample S can be represented by<F, D), where F is a feature vector composed of D random variables (F1, F2,..., FD), and D is the category to which S belongs. This way, a sample can be represented as a D+1 dimensional vector. The observation value of a sample is the value of the feature vector F. So, under the condition of a certain observation value (f1, f2,... fD), the probability of the current sample belonging to category c is equation (2):

$$p(c=ci， \mathrm{F1}=f1, F2=f2, Fn=fn) \tag{2}$$

For each classification, calculate the above probabilities separately, so that the classification with the highest result is used as the classification of the current sample. According to Bayesian theorem, as shown in equation (3):

$$p(c=ci， \mathrm{F1}=f1, F2=f2, Fn=fn)=\frac{p(\mathrm{F1}=f1, F2=f2, Fn=fn\,\mathrm{C}=\mathrm{Ci})}{p(\mathrm{F1}=f1, F2=f2, Fn=fn} \tag{3}$$

The Bayesian belief network improves the naive Bayesian algorithm by utilizing the conditional independence of random variables. Given three random variables X, Y, and Z, when X, Y, and Z satisfy the following equation, X and Y are independent of the Z condition, as shown in equation (4):

$$p(x,y,z)=p(x,z) \tag{4}$$

The Bayesian belief network consists of two parts. The first part is a directed acyclic G (V, E), where each point of G represents a random variable in the feature vector; An edge from point X to point Y indicates that the value of Y depends on X, and X is called the parent node of y. The set of all parent nodes of Y is denoted as parent (Y). BBN assumes that each random variable is conditionally independent of other nodes given all parent nodes. The other part is the Conditional Probability Table (CPT) for each random variable, which provides the conditional probabilities of the corresponding random variable with respect to all its parent nodes[7-8]. For variables without parent nodes, the conditional probability degenerates into a prior probability. Given the network structure and corresponding CPT of BBN, the probability of any sample appearing can be expressed as equation (5):

$$p(\mathrm{F1}=f1, F2=f2, Fn=fn， \mathrm{C}=\mathrm{ci}) \tag{5}$$

Constructing a Bayesian belief network through training data mainly involves two steps: first, determining the network structure, which is the dependency relationship between various random variables; The second is to calculate the CPT for each random variable. The network structure can be given by experience or determined through certain algorithms; When there are no hidden variables, the value of CPT can be obtained by counting the frequency of corresponding samples in the training data. When there are hidden variables, the value of CPT can be obtained through gradient training or EM algorithms.

C. *Deployment phase*

During the deployment phase, read the consumption records and common abnormal transaction records of users in the database to construct a Bayesian belief network. Mark the user's existing consumption records as normal, and mark common abnormal transaction records as abnormal. For each user, the k-means algorithm is used to discretize the values of transaction location and transaction amount.

As mentioned earlier, constructing a BBN network involves determining the network structure and calculating the CPT for each variable. Due to the fact that the network structure reflects the dependency relationships between various random variables, for ease of implementation, it is assumed that the dependency relationships of each random variable are the same for all user consumption transactions. In this way, our dependency analysis algorithm only needs to run once to construct a unified network structure for all consumption transaction records. For example, based on empirical knowledge, the following random variable orders are given: classification, transaction time, transaction location, merchant category, and transaction amount. Figure 2 shows a possible dependency relationship.

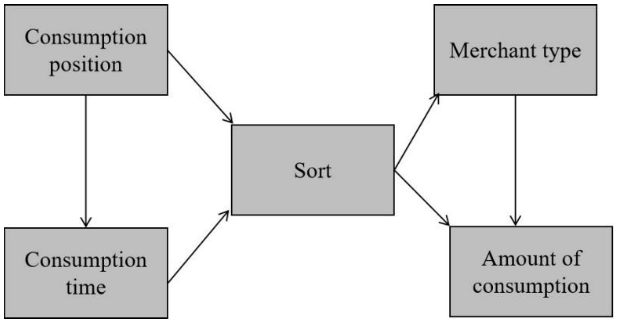

Figure 2. A possible dependency structure

D. *Testing stage*

When a financial consumption transaction is detected, the user ID and feature vector of the transaction are extracted and submitted to the BBN classifier. The classifier queries the corresponding CPT based on the user ID and calculates the posterior probability that the current transaction belongs to a normal transaction. This posterior probability is the trust factor TS at the business level of the transaction. If TS is less than a certain threshold, it marks the current transaction as abnormal at the business level[9-10]. Definition and calculation of trust factors TS and OS:

Business layer trust factor (TS): The confidence level output by a Bayesian belief network (BBN) model based on business layer data such as transaction amount, transaction frequency, geographic location, etc., represents the credibility of transactions at the business level.

Operational layer trust factor (OS): Based on operational layer data such as user behavior patterns, device fingerprints, login time, etc., the matching degree output by the BLAST-SSAHA model represents the credibility of transactions at the operational level.

IV. Final judgment

Based on TS and OS, we comprehensively judge the abnormal situation of the current transaction, and Table 1 lists our judgment method.

TABLE I. Final Judgment of Abnormal Situation

	Normal business level	**Abnormal business level**
Normal operation level	The transaction is normal	Whether the transaction is abnormal is determined by TS+OS.
Abnormal operation level	Whether the transaction is abnormal requires further manual audit.	Abnormal transaction

Firstly, it is worth noting that anomaly detection at the business level is targeted at the current single transaction, while anomaly detection at the operational level involves multiple operations. Therefore, anomalies at the operational level may or may not be related to whether the current transaction is abnormal. The handling of situations where both the business and operational levels are normal or abnormal is relatively simple, simply mark the corresponding transaction as normal or abnormal. For transactions that are abnormal at the business level but normal at the operational level, the value of TS+OS is used to determine whether the transaction is abnormal. The basis for doing this is that when a transaction is abnormal at the business level, its TS value is small. However, if the current operation sequence is very consistent with the normal operation sequence (with a large OS), then the current transaction is likely caused by unexpected events or short-term changes in consumer habits, and the transaction may still be classified as a normal transaction. It can be seen that this method reduces the false alarm rate of the detection method.

For transactions that are normal at the business level but abnormal at the operational level, we believe that additional manual audits are needed to determine whether they are abnormal transactions. Because the operations that cause abnormal operation sequences may be related to or unrelated to this transaction. If it is related to this transaction, it should be considered abnormal; Otherwise, it should be considered normal for this transaction, but other abnormal situations have been detected. It can be seen that this method reduces the false alarm rate of the detection method.

V. Conclusion

In this paper, we propose a method for detecting abnormal transactions in financial institutions such as banks. Our method mainly focuses on anomaly detection of user consumption transactions, and analyzes the current transaction at the business and operational levels through data mining. Finally, we comprehensively consider the analysis results at the business and operational levels to determine whether there are anomalies in the current transaction. Through analysis, it can be seen that combining two levels of analysis methods can more comprehensively discover abnormal situations in the payment system, which is conducive to reducing the false alarm rate and false alarm rate of detection, and improving the accuracy of abnormal transaction detection.

References

[1] Akhavan, M., & Hasheminejad, S. M. H. (2023). A two-phase gene selection method using anomaly detection and genetic algorithm for microarray data. Knowledge-Based Systems, 2(6)2, 110249-.

[2] Zhang, C., Zhao, Y., Zhou, Y., Zhang, X., & Li, T. (2022). A real-time abnormal operation pattern detection method for building energy systems based on association rule bases. Building Simulation, 15(1), 69-81.

[3] Zhou, Y., Zhang, C., Qin, M., Zhang, Y., & Wang, J. (2022). An improved top-eye trajectory anomaly detection method integrating semantic features. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 29(0)0-2903.

[4] Wen, L. I., Dejian, L. I., Yongyue, M. A., Wang, T., Xin, W., & Jie, L. I. (2024). Research on denoising of joint detection signal of water quality with multi-parameter based on ieemd. Optoelectronics Letters(2), 107-115.

[5] Mallampati, S. B., & Seetha, H. (2024). An integrated feature extraction based on principal components and deep auto encoder with extra tree for intrusion detection systems. Journal of Information & Knowledge Management, 23(01).

[6] HongmeiYAN, & MingyiHE. (2023). Hyperspectral data band selection based on clustering joint skewness-kurtosis index. Journal of Signal Processing, 39(1), 1-10.

[7] Wei, F. (2023). Study on behaviour anomaly detection method of english online learning based on feature extraction. International Journal of Reasoning-based Intelligent Systems, 15(1), 41-.

[8] Zhang, Z., Li, W., Ding, W., Zhang, L., Lu, Q., & Hu, P., et al. (2023). Stad-gan: unsupervised anomaly detection on multivariate time series with self-training generative adversarial networks. ACM transactions on knowledge discovery from data(5), 17.

[9] Aksu, D., & Aydin, M. (2022). Mga-ids: optimal feature subset selection for anomaly detection framework on in-vehicle networks-can bus based on genetic algorithm and intrusion detection approach. Comput. Secur., 1(1)8, 102717.

[10] Li, X., Huang, T., Cheng, K., Qiu, Z., & Sichao, T. (2022). Research on anomaly detection method of nuclear power plant operation state based on unsupervised deep generative model. Annals of nuclear energy65(Mar.), 167.