Research on Anomaly Detection Algorithms for Financial Data Based on Angle

Yuanyuan Hong*
Rongzhi College of Chongqing Technology and Business University, Chongqing 401320, China
*Correspondence author email: 707557056@qq.com

Abstract. With the emergence of more and more massive data, under the background of large data, the existing angle-based anomaly detection algorithms have the problem of too much computation. Based on this, this paper improves the angle-based anomaly detection method, and proposes an angle-based anomaly detection method based on Data Center for anomaly detection of large amount of network financial transaction data. By establishing data update mechanism for different data sets, real-time data detection can be carried out.

1. Introduction
Existing anomaly detection algorithms proposed at home and abroad include distribution-based, clustering, classification, depth, density and distance. Previous anomaly trading detection methods are mostly based on statistics. With the progress of new research fields and methods, more and better detection methods have been proposed. It includes anomaly detection method based on artificial neural network, anomaly recognition using partition method, anomaly detection using fuzzy rough set and anomaly detection algorithm based on self-organizing mapping technology.

With the maturity of technology, the number and latitude of data have increased by leaps and bounds. When anomaly detection is carried out against the background of large data of TB level and 100 kilo latitude, the traditional anomaly detection algorithm needs nearest neighbor search, which makes the detection results based on distance or density difficult to recognition and operation time is unbearable. When existing data mining methods detect high-dimensional data, their computational complexity is mostly $O(n^2)$ ($n$ is the number of data objects). When calculating the distance of full-dimensional space in high-dimensional case, the detection effect is very different and the error rate is high. Therefore, the current research focuses on how to improve the measurement effectiveness and computational efficiency of anomaly detection algorithm.

In order to solve the problem of "data dimension disaster" in anomaly detection, Kriegel proposed a method based on angle distribution to calculate potential anomaly points in high-dimensional datasets. On this basis, Pham et al. put forward the formula of anomaly factor based on angle distribution. In high-dimensional data space, the angle is more stable than the usual distance, and the method based on angle distribution will not make the data disaster worse substantially. However, with the emergence of more and more massive data, under the background of large data, the existing angle-based anomaly detection algorithms such as FastABOD, FastVOA and so on have the problem of too much computation. To solve this problem, a new outlier detection (C-ABOD) algorithm for high-dimensional data based on angle distribution is proposed. At the same time, real-time data detection can be achieved by establishing data update mechanism for different data sets.
2. Anomaly detection thought based on angle variance

Kriegel proposed an anomaly detection method based on angular distribution, which judges the anomaly points by calculating the anomaly factors of each data element. The basic ABOD algorithm has a very high precision, but it also has the problem of too much calculation. It is a very time-consuming work to find outliers in large data sets. For the outlier detection in the background of massive data, Kriegel and others have proposed improved FastABOD algorithm and LB-ABOD algorithm.

FastABOD algorithm uses the angular variance of the nearest neighbor of a point as an approximation of the outlier factor of the angular variance of a point in the data set. This method effectively speeds up the operation efficiency of ABOD algorithm. However, since the running time of the algorithm and the detection results depend on the number of the nearest data points, and need to be specified by the user, different values will get different results, so the selection of values is a relatively difficult problem. At the same time, with the increase of data size and data dimension, the running time and results of the algorithm will become worse. The improved LB-ABOD algorithm adds error values to correct the operation results so as to achieve good results in the case of massive data and high dimensions. There is no essential difference between FastABOD algorithm and LB-ABOD algorithm in computing time. The main difference between the two algorithms is that the calculation of error value is added in LB-ABOD algorithm to obtain higher accuracy, but when calculating error value, it will consume a certain amount of memory space and CPU.

Based on the idea of angle variance, another angle-based anomaly detection algorithm FastVOA is proposed in the literature. The angle cosine variance weighted by the corresponding distance of the data points is used as the anomaly factor to estimate the degree of anomaly. The algorithm has approximate linear time, but it has the problem of too many user-defined parameters. Because different parameters are needed, the detection results are affected.

3. C-ABOD algorithm

Aiming at the problems of the above algorithms, a new improved ABOD algorithm called C-ABOD algorithm is proposed. Compared with FastVOA and FastABOD algorithms, C-ABOD algorithm reduces the computational overhead effectively, speeds up the computational speed of the algorithm, reduces the time and physical storage requirements of the algorithm. At the same time, the algorithm has a good accuracy. On the premise of guaranteeing the detection effect, the running time is shorter and the overhead is smaller.

The angular variance of data points is defined as follows:

Definition 1: Given point set \( S \subseteq \mathbb{R}^d \), point \( A \in S \). Definition norm: \( ||\cdot||: \mathbb{R}^d \rightarrow \mathbb{R}_+^d \). The inner product is defined as \( \langle \cdot, \cdot \rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \). For two points \( B \in \mathbb{R}^d \) and \( C \in \mathbb{R}^d \) in \( S \), \( \overrightarrow{BC} \) denotes vector \( \overrightarrow{AB} \) and \( \overrightarrow{AC} \). The abnormal factor \( \text{VAR}_{A,B,C}(\mathbf{A}) = \frac{\left(\mathbf{A} \cdot \mathbf{B}\right)}{\left(\mathbf{A} \cdot \mathbf{C}\right)} \) of the conclusion angle is the variance of the angle.

Angle-based anomaly factor ABOF (A) is the sum of angular variance with distance as the weighting factor, which is composed of the remaining two points in point \( A \) and set \( S \).

\[
\text{ABOF}(A) = \text{VAR}_{A,B,C}(\mathbf{A}) = \sum_{B \in S \setminus \{A\}} \left(\frac{\mathbf{A} \cdot \mathbf{B}}{\mathbf{A} \cdot \mathbf{C}}\right)^2 = \frac{\sum_{B \in S \setminus \{A\}} \left(\frac{\mathbf{A} \cdot \mathbf{B}}{\mathbf{A} \cdot \mathbf{C}}\right)^2}{\sum_{B \in S \setminus \{A\}} \left(\frac{1}{\mathbf{A} \cdot \mathbf{C}}\right)^2} = \sum_{B \in S \setminus \{A\}} \left(\frac{\mathbf{A} \cdot \mathbf{B}}{\mathbf{A} \cdot \mathbf{C}}\right)^2 - \left(\frac{1}{\mathbf{A} \cdot \mathbf{C}}\right)^2
\]  \( (1) \)
Take any point A in point set $S$, take A as the vertex, and choose the other two points B and C in point set $S$ to form vectors $\overrightarrow{AB}$ and $\overrightarrow{AC}$ where $B \in S \setminus \{A\}, C \in S \setminus (A, B)$. By calculating the angular variance of vectors $\overrightarrow{AB}$ and $\overrightarrow{AC}$, as the ABOF value of the angle-based anomaly factor of data point A, the angular variance of all data points is ranked from small to large. The angular variance is regarded as the anomaly factor, and the data point in the front ranking is the anomaly data.

Concrete calculation process of C-ABOD algorithm is as follows:

1) Data Normalization Processing

Firstly, the dimension of data is normalized, which can reduce the weight of calculation between different dimensions. In order to eliminate the influence of dimension and the influence of variable variation and numerical value on outlier detection, it is necessary to standardize the attributes on the data set before calculating. In the data normalization stage, users can also assign different weights to different dimension data according to their actual needs to meet the actual needs.

2) Selecting Data Center Points

Existing anomaly detection algorithms are based on absolute data values, without considering the change of data coordinate system. As shown in Fig. 1, the selection of coordinate system origin $O$ and $O'$ is different, which makes the data points relative to coordinate system origin change correspondingly, but the relative positions between data points do not change their properties, which provides a new reference method for the detection of data outliers. Using the method shown in Figure 1, the virtual transformation of the coordinate system of data points can be carried out. By utilizing the nature of the variant data set itself, the central point of the data set can be calculated and used as the origin of the new coordinate system of the data set. The outlier detection can be carried out by utilizing the nature of the data set itself.

![Figure 1. Virtual coordinate coefficient data center $O'$](image)

Calculate the central point of the data set. Firstly, the $3\sigma$ principle is used to preprocess the data, eliminating the influence of individual large or small data on the central point calculation. This is because if there is too large or too small data in the data, when calculating the center point, it will make the center point move towards the direction of too large or too small data, so that the calculated center point will deviate from the actual center point, making the calculation value of anomaly factor inaccurate. After that, the average values of all dimensions of data are calculated, and the average values of each dimension are obtained to form a new data center point $O'$.

3) Angular variance calculation

In this paper, such anomaly factors are also used to estimate the anomaly degree of data points. Every data in the data set is calculated for the angle variance of the anomaly factor. The smaller the value of the angle variance anomaly factor is, the greater the possibility of the anomaly point is.
FastABOD algorithm uses the angular variance of points A’s k nearest neighbor data points as the approximation of the angular variance anomaly factor of point A in data set S. This method effectively speeds up the operation efficiency of ABOD algorithm. Therefore, in the C-ABOD algorithm, the angular variance of k nearest neighbor data points of point A is also used as the approximate estimation of the global angular variance of data points.

In order to improve the accuracy of angle variance estimation of k nearest neighbor data points, an error value is added to the improved LB-ABOD algorithm to correct the operation results, so that good results can be obtained in the case of massive data and high dimensions. Therefore, combined with the above outlier detection algorithm, the following angular variance calculation method is adopted in this paper, as shown in formula (2).

\[
C - ABOF(A) = \text{VAR} \left( \frac{\langle AB, AO \rangle}{\|AB\| \ast \|AO\|^2} \right) = \sum_{b \in N_k(A)} \left( \frac{\langle AB, AO \rangle}{\|AB\|^2 \ast \|AO\|^2} \right)^2 + R^2
\]

The new point O is the center point of all data points. It is a new point composed of the average value of each dimension. Residual R1 and R2 are used as approximate estimates of the error values of the angle variance anomaly factor of point A in set Nk(A) and point A in set S.

Calculation of R1 and R2 is the same as that of R1 and R2 in LB-ABOD algorithm. The smaller of A is the better of B. The calculation of G is as follows:

\[
R^2 = \sum_{b \in B} \left( \frac{1}{\|AB\|^3 \ast \|AO\|^3} \right) - \sum_{b \in N_k(A)} \left( \frac{1}{\|AB\|^3 \ast \|AO\|^3} \right)
\]

According to formula (2), the angular variance of all points in the data set is calculated and used as the anomaly factor of the data points. The smaller the data value of the angular variance anomaly factor, the greater the possibility of the anomaly point.

When calculating with C-ABOD algorithm, by means of coordinate transformation and using the nature of data itself, the central point of the data set is selected, and the angle common edge through which the central point is fixed as the common vertex is selected, which effectively reduces the computational overhead. However, the C-ABOD algorithm also has some limitations. For some special data points distribution, there is no abnormal point detection. When using the C-ABOD algorithm, the angle variance of the abnormal factor of the abnormal point is also very large. At this time, the C-ABOD algorithm can not detect such abnormal points. It is necessary to adjust the algorithm according to the special situation. When accelerating the algorithm, the density-based neighborhood selection is used instead of the distance-based neighborhood selection.
4. Experiment

In order to verify the performance of this algorithm, two aspects of scalability of data dimension and data scale are used as evaluation indicators. Other algorithms such as ABOD algorithm, FastABOD algorithm, FastVOA algorithm and MLOF algorithm are selected as reference comparisons to compare the artificial data sets and real data sets.

In order to reduce the error of detection results, this paper runs all the algorithms 50 times to find their average value, and compares the performance of each algorithm in terms of precision and running time. Because FastABOD algorithm and C-ABOD algorithm as well as MLOF algorithm need to set parameters, in this experiment, because 10 outliers are selected, set \( k = 10 \).

At the same time, the UCI machine learning library is selected as the actual data set of the classification and machine learning tasks, and Isolet is used as the anomaly detection algorithm. The data set includes the pronunciation data of 26 letters of the alphabet. In addition, the real customer transaction data set of a bank in one month was selected for experiment, which contains more than 120,000 transaction records of more than 10,000 customers.

4.1. Scalability of data dimension

In order to test the scalability of the data dimension of the algorithm, the dimension parameters are set to 25, 50 and 100 using simulated data sets. The operation of ABOD algorithm, FastABOD algorithm, C-ABOD algorithm and MLOF algorithm are analyzed.

Experiments show that C-ABOD algorithm performs the best, regardless of the data dimension, it has a good calculation results, the accuracy has been maintained at 1.0; ABOD algorithm and MLOF algorithm with the increase of dimension, the accuracy has also increased, and maintained at 1.0 in high-dimensional data; FastABOD algorithm with the increase of dimension. Additionally, the accuracy is increased, but lower than the other three algorithms. From the point of view of operation time, with the increase of dimension, the operation time of the four algorithms increases, while the operation time of MLOF algorithm is less, and with the increase of dimension, the growth rate of C-ABOD algorithm is also smaller. The growth rate of C-ABOD algorithm is also significantly less than that of FastABOD algorithm, while the running time of ABOD algorithm is from increment to increment. On the other hand, they are significantly inferior to the three algorithms mentioned above.

4.2. Scalability of data size

With the increase of data scale, the accuracy of C-ABOD algorithm has been kept at about 1.0, which has good stability. With the increase of data scale, the accuracy of MLOF algorithm and ABOD algorithm increases, but in low dimensions, the accuracy decreases with the increase of data scale; whereas the accuracy of FastABOD algorithm is ignored. The accuracy of low-dimensional data or high-dimensional data decreases with the increase of data size. In addition, from the point of view of running time, the running time of the four algorithms increases with the increase of data scale. MLOF algorithm performs best, and C-ABOD algorithm takes less time than FastABOD algorithm. Because of the long running time of ABOD algorithm, it is obviously inferior to the two algorithms in terms of running time increment and speed-up. The experimental results show that the C-ABOD algorithm has better data scalability than other algorithms.

4.3. The Effect of \( k \) Value Selection on the Result

Because C-ABOD algorithm, FastABOD algorithm and MLOF algorithm all need to determine the nearest neighborhood, so different values will have an impact on the operation results. When selecting 50 dimensions, the data set with the size of 3000 is selected, and \( k \) is taken 5, 10, 15, 20, 25 and 30, respectively, to study the influence of different values on the operation results.

C-ABOD algorithm and MLOF algorithm have good accuracy in any case, while FastABOD algorithm's operation accuracy increases with the increase of value, and all algorithms'operation time increases with the increase of \( k \) value. Among them, C-ABOD algorithm and MLOF's operation time increases slowly, but it can be seen that C-ABOD algorithm's operation time increases slowly. The time
increment of FastABOD algorithm is obviously less than that of MLOF algorithm, and the operation time of FastABOD algorithm is the highest, and its running time increment rate is obviously greater than that of the other two algorithms. Through the above analysis, we can see that FastABOD algorithm has the longest operation time and higher accuracy, while C-ABOD algorithm and MLOF algorithm have good detection effect regardless of the size of the value, and have better performance than FastABOD algorithm. MLOF algorithm runs less time than C-ABOD algorithm. Experiments on Isolet outlier detection in the following real data sets show that C-ABOD algorithm is superior to MLOF algorithm.

4.4. Representation of the algorithm in real data sets

Fig. 2 (a) shows the accuracy of the three algorithms for Isolet outliers in real data sets. Figure 2 (b) is the time required for the three algorithms when the accuracy is 1.0. Because of the long operation time of FastVOA algorithm, only part of the values is shown in the figure. In outlier detection of real data set Isolet, the detection accuracy when \( k = 10 \) and the values of each algorithm when the accuracy is 1.0 are calculated respectively. The experimental results show that the accuracy of FastABOD algorithm, MLOF algorithm and FastVOA algorithm is 0.0, while the accuracy of C-ABOD algorithm is 1.0, and the performance is perfect. Further experimental results show that even when \( k = 1 \) is selected, the C-ABOD algorithm can achieve a good accuracy of 1.0. The 10 top-level data points just contain all 10 outliers. When \( k = 14 \), FastABOD algorithm also achieves accuracy of 1.0; MLOF algorithm achieves accuracy of 1.0 only when \( k = 30 \); and FastVOA algorithm chooses \( t = 10, S1 = 10, S2 = 10 \) because of its many parameters, the accuracy is only 0.5, which is the worst performance. When the accuracy is 1.0, the operation time of C-ABOD algorithm is slightly better than that of MLOF algorithm, and the performance of C-ABOD algorithm is the best.

In addition, the real customer transaction data set of a bank in one month was selected for experiment, which contains more than 120,000 transaction records of more than 10,000 customers. During the experiment, the selected data set is randomly divided into two parts. 70% of the data is randomly selected as training data and the remaining 30% as test data. Through training data, the parameters of the algorithm can be learned. When the test data is changed, some transaction records are randomly selected for each change, and then modified by adding perturbations to the transaction records. When adding perturbations, it is divided into two steps. The first step is to make only minor changes to some data, which is called perturbation data, so that the modified transaction records are more original than the original ones. The transaction record has only slightly changed, but it is not enough to be an abnormal transaction record. The second step is to change some transaction record data and mark it as abnormal transaction. In addition, considering the generalization of the algorithm, different transaction records are randomly selected for processing each time. Each experiment is repeated 50 times, and its average value
is taken as the final test result, in order to reduce the impact of different data sets on the detection results of the algorithm and enhance the scalability of the algorithm. The results are shown in Figure 3. As can be seen from Figure 3 (a), the C-ABOD algorithm has the best detection results under different values, and the accuracy of the other three algorithms varies with the different values. In addition, it can be seen from Figure 3 (b) that the C-ABOD algorithm has the minimum operation time.

Figure 3. Results of the calculation of the bank's real transaction data

Comparing C-ABOD algorithm, ABOD algorithm, FastABOD algorithm and MLOF algorithm in four aspects: scalability of data dimension and scalability of scale, the experimental results show that: in the same dimension, with the increase of data point size, the best performance is C-ABOD algorithm; in different dimensions, with the increase of data dimension, the best performance is C-ABOD algorithm. With the increase of points, MLOF is the fastest algorithm, followed by C-ABOD algorithm. The experimental results show that C-ABOD algorithm performs the best. It has good calculation results no matter in any data dimension. The accuracy has been maintained at 1.0, and has good stability. For selecting different values, the experimental results show that for No. At the same value, C-ABOD performs well in both accuracy and computation time. According to the comprehensive evaluation, the C-ABOD algorithm has the strongest superiority.

5. Conclusion

This paper improves the existing angle-based anomaly detection method to solve the disaster problem of high-dimensional spatial dimension, and proposes an angle-based anomaly detection method based on Data Center for anomaly detection of large amount of network financial transaction data. By establishing data update mechanism for different data sets, real-time data detection can be carried out. Through the collection of historical transaction data, it can accurately detect abnormal transactions of users and reduce the occurrence of missed detection and misjudgement. The collection and analysis of abnormal transaction pattern data has been recognized as the core issue of combating money laundering crime. In addition, we can effectively monitor and supervise the abnormal transactions by analyzing and mining the abnormal transaction behaviors such as the amount, frequency, source, direction and use of transactions in the network transactions.

References

[1] Ding Z, Fei M. An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window[J]. IFAC Proceedings Volumes, 2013, 46(20):12-17.
[2] Dansereau C L, Pierre B, Kangjoo L, et al. Detection of abnormal resting-state networks in individual patients suffering from focal epilepsy: an initial step toward individual connectivity assessment[J]. Frontiers in Neuroscience, 2014, 8.
[3] Whelton N Q, Killen L, Buckley F. Detection of Abnormal Recordings in Irish Milk Recorded
Data[J]. Irish Journal of Agricultural & Food Research, 2007, 46(2):117-127.

[4] Shin J, Lee B, Suk Park K. Detection of Abnormal Living Patterns for Elderly Living Alone Using Support Vector Data Description[J]. IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society, 2011, 15(3):438.

[5] Yi T H, Guo Q, Li H N. The research on detection methods of GPS abnormal monitoring data based on control chart[J]. Engineering Mechanics, 2013, 30(8):133-141.

[6] Ding J, Liu Y, Zhang L, et al. An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model[J]. Applied Intelligence, 2016, 44(2):340-361.

[7] Rassam M A, Maarof M A, Zainal A. Adaptive and online data anomaly detection for wireless sensor systems [J]. Knowledge-Based Systems, 2014, 60:44-57.

[8] Zhu W K, Li H F, Mu J S, et al. An Anomaly Detection and Correction Method Based on Measurement Representation Model[J]. Applied Mechanics and Materials, 2013, 333-335(1):51-57.

[9] Yin C, Zhang S, Yin Z, et al. Anomaly detection model based on data stream clustering [J]. Cluster Computing, 2017.

[10] H. H. P, G. D, S. H. Two-tier network anomaly detection model: a machine learning approach [J]. Journal of Intelligent Information Systems, 2015, 48(1):1-14.