

## *Retraction*

# **Retracted: False Financial Statement Identification Based on Fuzzy C-Means Algorithm**

### **Complexity**

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

Copyright © 2024 Complexity. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

### **References**

- [1] J. Li, "False Financial Statement Identification Based on Fuzzy C-Means Algorithm," *Complexity*, vol. 2021, Article ID 5522894, 11 pages, 2021.

## Research Article

# False Financial Statement Identification Based on Fuzzy C-Means Algorithm

Jixiao Li 

Asset Management Office, Shandong University of Arts, Jinan 250300, China

Correspondence should be addressed to Jixiao Li; [z00799@sdca.edu.cn](mailto:z00799@sdca.edu.cn)

Received 26 February 2021; Revised 19 March 2021; Accepted 23 March 2021; Published 31 March 2021

Academic Editor: Zhihan Lv

Copyright © 2021 Jixiao Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Financial accountants falsify financial statements by means of financial techniques such as financial practices and financial standards, and when compared with conventional financial data, it is found that the falsified financial data often lack correlation or even contradict each other in terms of financial data indicators. At the same time, there are also inherent differences in reporting patterns from conventional financial data, but these differences are difficult to test manually. In this paper, the fuzzy C-means (FCM) clustering method is used to amplify these differences and thus identify false financial statements. In the proposed algorithm, firstly, the normalization constraint of the FCM clustering algorithm on the sum of affiliation of individual samples is relaxed to the constraint on the sum of affiliation of all samples, thus reducing the sensitivity of the algorithm to noise and isolated points; secondly, a new affiliation correction method is proposed to address the problem that the difference in affiliation is too large after the relaxation of the constraint. In the discussion of this paper, most of the information comes from the annual reports of companies, administrative penalty decisions of the Securities Regulatory Commission, and some information comes from research reports made by securities institutions, which are limited sources of information. The proposed method can correct the affiliation to a reasonable range, effectively avoiding the problem that some samples have too much affiliation and become a class of their own and also avoiding the problem that it is difficult to choose the termination threshold of the algorithm iteration due to too little affiliation, and can ensure that the constraint on the sum of affiliation of all samples is always satisfied during the iteration of the algorithm. The method has the characteristics of high recognition accuracy and has the significance of theoretical method innovation.

## 1. Introduction

Economic globalization is the development trend of today's world economy, accounting as a business language is moving towards the international process, financial fraud has also become a major persistent problem prevailing in countries all over the world, and the manifestations of financial fraud have shown hidden and diversified characteristics [1]. The emergence of these problems, on the one hand, disrupts the normal order of the securities market and affects the healthy development of the securities market; on the other hand, it also poses a challenge to the supervision of securities and intermediaries. For investors, distorted financial reports of listed companies will prevent them from understanding the true, complete, and accurate actual

situation of listed companies, resulting in disinvestment and great economic loss; for the company itself, once the matter is revealed, it will greatly damage the brand effect of the company and destroy the image of the company; when financial fraud becomes more and more rampant, it will prevent the development of the industry and the government will lose its credibility [2]. The government will lose its credibility and will hinder the healthy and orderly development of the securities market. Therefore, if we want a healthy and good securities market environment, we should have a set of checks and balances system to identify, supervise, and defend against counterfeiting, in order to stop and eliminate financial counterfeiting; for this reason, the study of financial counterfeiting has a more far-reaching theoretical and practical significance [3].

Through the analysis and research on the meaning, characteristics, and means of financial fraud, this paper can improve the discernment of financial fraud and lay the argumentative foundation for the research on the prevention of financial fraud. Through the in-depth analysis and research on financial fraud, the problems of financial fraud in practice are summarized, which plays a great theoretical significance for future research on the prevention of financial fraud and the forward development of the securities market [4]. Through the study of financial fraud, governance strategy can promote the improvement of the company's internal control, promote the improvement of auditing standards and the revision of relevant laws and regulations, provide a reference for the construction of financial fraud strategies and methods, help to stifle the company's financial fraud, improve investors' ability to identify fraudulent behavior, inhibit the prevention of intermediaries to provide help for the company's financial fraud, strengthen the supervision of relevant government departments, and improve. The paper adopts a combination of normative and empirical approaches.

The recognition of false financial statements based on the fuzzy C-means (FCM) algorithm checks the false financial statements and reduces the incidence of false financial statements. This paper adopts a combination of normative and empirical methods to identify the phenomenon of corporate financial fraud. The empirical study focuses on establishing a logistic regression model and testing it. The research hypothesis establishes a model which can identify financial fraud of companies and test the prediction sample. The model effect is compared with the financial fraud identification effect of other companies, and finally, the model is tested again with the latest financial fraud cases. This paper is divided into four parts: the first part is the introduction, which introduces the research background and significance of this paper and introduces the research ideas of this paper. The second part is the study of false financial statement identification based on FCM algorithm, reviewing and evaluating the current situation of financial falsification at home and abroad, the determination of financial falsification model samples and indicators, analyzing the problem that the basic FCM algorithm is sensitive to noise and isolated points, then analyzing three related algorithms for affiliation correction, and explaining the defects of the algorithms. And finally, a brief description of other algorithms for affiliation correction is given [5]. Firstly, the research objective of this paper is determined, and then a sample financial fraud model is established. Then, based on the symptoms, characteristics, and causes of fraud, the indicators that are used to build the model are identified. The third part is the testing and application of model validity. First, the model is substituted into the prediction sample for testing, and then the model is substituted into other falsifying companies for comparative testing. The fourth part is the conclusion and recommendations. The previous studies are summarized and analyzed, recommendations are made,

the main contributions of this paper are noted, and limitations and prospects for future research are pointed out.

## 2. Research on False Financial Statement Identification Based on FCM Algorithm

*2.1. Related Work.* Foreign research related to financial fraud started earlier and the theoretical research on financial fraud is also more systematic, and the classical fraudulent motive theory and fraudulent behavior theory formed are widely adopted by national research. Kaja et al. proposed a new fraud triangle theory consisting of motive, opportunity, and integrity, based on criminologist Cressey's fraud model [6]. Sheeba et al. modified the fraud diamond theory by adding the "arrogance" factor to the Pentagon fraud theory, which identifies five elements of financial fraud: pressure, opportunity, rationalization (excuses), competence, and arrogance [7]. Saeed et al. categorized the major financial frauds as inflated revenues or assets and deliberate concealment or nondisclosure of material events [8]. Singh et al. found that inflating revenues or assets appears to be more common when financial fraud is involved. The means of financial fraud, which is the key to commit fraud, have become increasingly sophisticated, diverse, and insidious [9]. Existing research shows that the most common methods of falsification today are fictitious revenue and fictitious assets; the flexibility of accounting standards and the specificity of related party transactions account for the majority of falsifications. In recent years, the more secretive methods involve "cash flow" supplemented by "information flow" and "logistics," in which fictitious funds are usually created with the help of accounts of related parties, suppliers, and customers and transferred to asset accounts such as inventory and fixed assets, so that the falsification forms a complete closed loop.

The FCM algorithm is originally designed to cluster the set of points in the space. In practice, there are also incomplete data to deal with, as well as other forms of data: mixed data and interval data, which cannot be directly clustered by the basic FCM algorithm. Therefore, more improved algorithms are proposed for clustering different forms of datasets. Nawaz and Yan proposed a nearest neighbor interval-based clustering method for incomplete data, based on the nearest neighbor principle to find the nearest neighbor's perfect samples based on incomplete samples, and used the nearest neighbor interval to represent incomplete data [10]. Arshad et al. applied the improved BP neural network to predict the missing attribute values of incomplete data and then used the FCM algorithm to cluster the perfect data [11]. For hybrid data, Yuan et al. weighted different types of attributes of hybrid data based on intra- and interclass information entropy to improve the clustering effect [12]. Zhang et al. extended the Euclidean distance and applied it to the clustering of hybrid data to measure the dissimilarity between samples and classes and improve the

clustering performance of the algorithm for hybrid data [13]. Zhao et al. proposed a similarity measure for hybrid data that fully consider the data information and reduce the influence of noise on the clustering results [14]. In the objective function of the FCM algorithm, an improved penalty term is introduced to enhance the effect of samples on the nearest clustering center under the affiliation normalization condition constraint and improve the clustering effect.

This paper studies the problem of membership correction in the FCM algorithm, proposes a reasonable correction method for membership based on relaxing the normalization constraints of membership, which reduces the algorithm's sensitivity to noise and outliers to a certain extent, and applies the proposed algorithm to the identification of false financial statements to verify the performance of the algorithm. On the basis of summarizing the results of previous studies, this paper adopts the method of case analysis, starting with the method of financial fraud in the analysis and looking for loopholes in the financial logic. The theory of financial fraud signals characteristics is a further extension. From the perspective of financial information, it explains that, in order to achieve the purpose of fraud, relevant information clues will be reflected in financial data. Therefore, this paper will immediately analyze the characteristics of financial fraud signals from the characteristics of financial data. Finally, the financial fraud risk factor is a comprehensive analysis and study of the entire financial fraud process [15]. It comprehensively analyzes the whole process of fraud from the subjective factors of financial fraud, objective conditions, environmental factors, and benefit cost measurement. The financial fraud factor is a summary of the fraud situation.

*2.2. False Financial Statement Factor.* The four major motivation theories of fraud are developed in sequence. Compared with other theories, the GONE theory has a higher degree of detail and a wider range and has been widely used at home and abroad. Although the iceberg theory and the fraud triangle theory generally summarize the motivations of financial fraud, the GONE theory puts forward the exposure factor, which has been further improved in theory, and combined with the increasingly complex status of my country's financial fraud, the GONE theory can also be seen. The application is more reasonable and can more deeply explore the motivation behind financial fraud [16]. This is also an important reason why this paper is selected from the GONE theory as the theoretical basis. Relatively speaking, the fraud risk factor theory has the widest coverage, but the motivation factor in this theory is not much different from the four-factor connotation in the GONE theory, and the GONE theory divides fraud motivation into external and internal causes, which is more beneficial. In the case analysis, distinguish the inside and outside and sort out the logic.

Generally speaking, the fraud triangle theory, the GONE theory, and the fraud risk factor theory are more

suitable for case analysis of financial fraud. Table 1 shows the difference and connection between the fraud triangle theory, the GONE theory, and the fraud risk factor theory.

The price of fraud being discovered is the illegal cost of financial fraud, which depends on the degree of perfection of laws and regulations and the intensity of punishment. Under normal circumstances, the higher the degree of punishment after fraud is discovered, the smaller the possibility of fraud is. Huge profit temptation is often the biggest motivation for companies to implement financial fraud, and the cost after fraud is discovered makes the company choose to weigh the pros and cons before implementing financial fraud. If the penalty for financial fraud is small, the company will choose to take the risk. Therefore, increasing the illegal cost of financial fraud can effectively curb the occurrence of fraud. Figure 1 is a diagram of the formation mechanism of financial fraud.

Fraud motives refer to the incentives and intentions of fraud. Companies in different industries and at different stages of development will have different motives for fraud. The three motives of financial fraud are the greed motive of stakeholders, the political pressure of local governments, and the requirement to maintain the leading position in the industry. Whitewashing financial statements through simple abnormal financial means can attract the attention of investors in the capital market and bring huge benefits to the company. In the face of this huge interest temptation, no matter the company's governance or other company stakeholders, there is undoubtedly sufficient motivation to cheat. From the perspective of individual management, the two company leaders who were punished used their operating control and information advantages to seek personal gains, instructed the financial department and the audit department to inflate business income, and resorted to various means to falsify and seek benefits. It seems that the company has been used as a personal profit-making tool, regardless of relevant laws and regulations, and greed has enabled the company and related individuals to obtain huge benefits in the process of financial fraud. However, from the perspective of a company, investors invest in a listed company because they know that the company will help the company realize value-added through various means, so that investors can get a lot of investment income.

*2.3. CM Algorithm False Identification Model.* Because the FCM algorithm only sets the normalization constraint on the membership degree of a single sample, the noise data far away from the center of each subcategory still has a large membership degree, which makes the algorithm sensitive to noise and outliers [17–20]. In contrast, IFP-FCM and Graphics Interchange Format (GIFP) FCM, which are based on affiliation correction, introduce different affiliation penalty terms in the clustering objective function to achieve the correction of affiliation of a single sample to different subclasses, which improves the clustering performance of the algorithm to some extent, but both algorithms are still constrained by the affiliation normalization condition in the

TABLE 1: Differences and connections between the three kinds of fraud motivation theories.

Theory type	Risk factor			
Fraud triangle	Pressure	Excuse	Opportunity	
GONE theory	Greedy	Need	Opportunity	Exposed
Fraud risk factor theory	Moral character	Motivation	Opportunity	Exposure
				Punishment

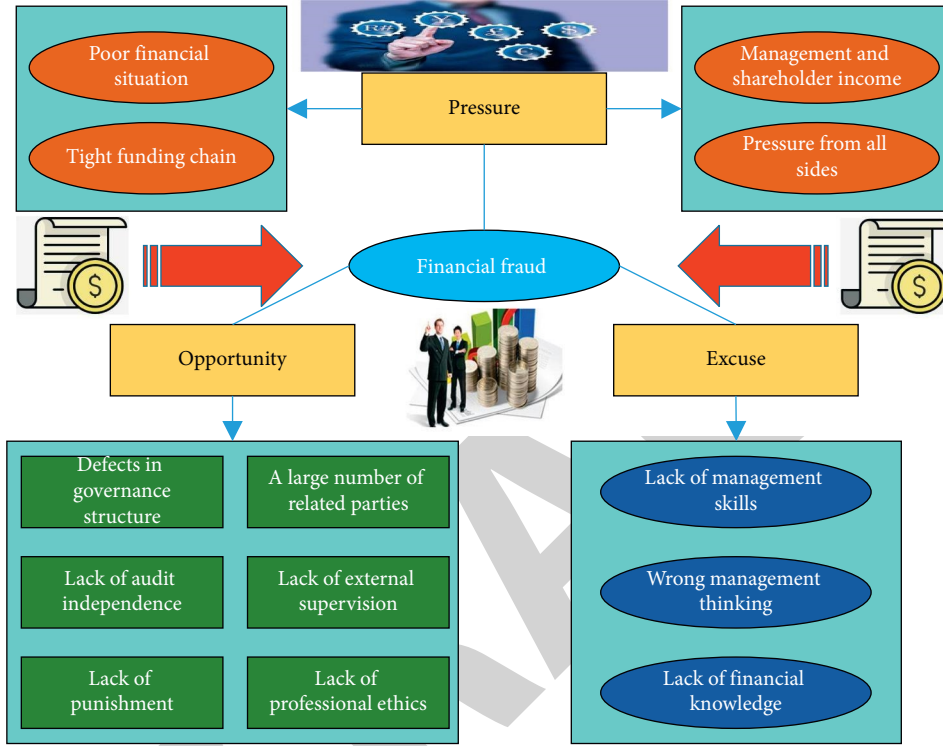


FIGURE 1: The mechanism of financial fraud formation.

clustering iteration process [21–23]. Given that affiliation can be interpreted as the force of samples on subclass centers, the normalization constraint of affiliation is equivalent to limiting the overall force of all samples on subclass centers in the data set to be the same, which affects the clustering performance of the algorithm to some extent. The flow of the FCM algorithm is shown in Figure 2.

Since the FCM algorithm sets the same constraint on the affiliation of each sample in the clustering process, which causes the algorithm to be sensitive to noise and isolated points, this paper relaxes the normalization constraint on the sum of the affiliation of individual samples to the constraint on the sum of the affiliation of all samples.

$$G(M, N) = \sum_{i=1}^a \sum_{j=1}^b m_{ij}^c |f(x_j) - n_i|^3. \quad (1)$$

The constraint on individual sample affiliation in the FCM algorithm is shown in equation (2).

$$\sum_{i=1}^a m_{ij}^c = \sigma * 100\%, \quad (j \subseteq [1, b]). \quad (2)$$

The constraint on the sum of affiliation of all samples is shown in equation (3).

$$\sum_{i=1}^a \sum_{j=1}^b m_{ij} = \sigma * b. \quad (3)$$

Given that the affiliation degree can be understood as a characterization of the attractiveness of a sample to a subclass center, the affiliation degree correction method proposed in this paper still accounts for a larger proportion of the attractiveness of the samples to the subclass center that can be guaranteed, while it can be controlled within a reasonable range to avoid the problem that these samples are too attractive to a subclass center leading to their self-contained class. And the affiliation value of the samples with the smaller sum of affiliation is kept constant, which can keep the lower attractiveness of these samples to each subclass center and reduce the sensitivity of the algorithm to noise and isolated points.

$$m_{ij} = m_{ik} * \frac{m_k^j}{m_k} * \frac{m_i^j}{m_i}. \quad (4)$$

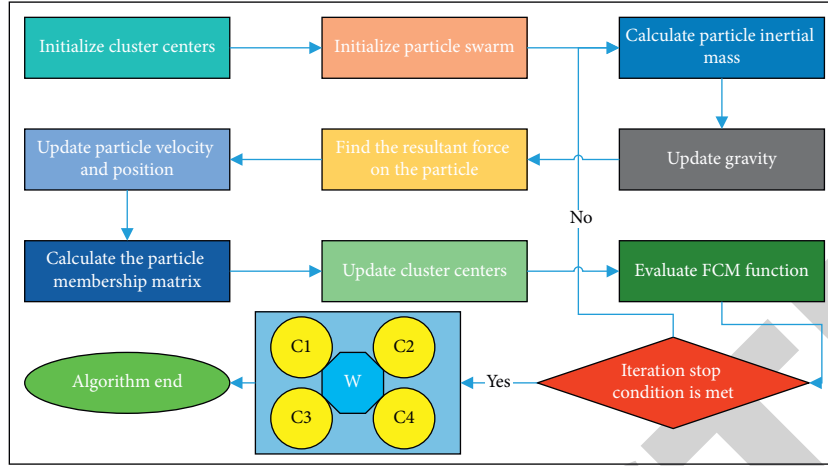


FIGURE 2: The FCM algorithm flow.

The affiliation function corresponds to the concept of “weight function” in robust statistics and the weight corresponding to the concept of “scale” can be obtained by calculation. Therefore, robust statistics can be used to estimate the weight of scales in the presence of noise. The clustering algorithm produces partitions, which are sets of  $m$  label vectors. For nonclear partitions, the usual method is applied to each column  $M_j$  of the matrix  $M$ .

$$N_j^K = K(M_j) = e_i \oplus m_{ij} > m_{jk}, \quad i, j, \exists k. \quad (5)$$

Updating the clustering centers using equation (5),

$$\exists i, j: \max \left( \frac{n_{ij} - m_j}{m_{ij} - n_i} \right)_2^{\sigma-1} \leq \omega. \quad (6)$$

The number of classes or clusters is usually denoted by  $G$ . If and when the algorithm is able to minimize the error function and utilizes fuzzy techniques or simple fuzzy in such algorithm, it is called the FCM method that uses fuzzy affiliation and assigns an affiliation to each data to a certain class. The importance of affiliation in fuzzy clustering is similar to that of pixel probability in the mixture modeling assumption. The benefit of FCM is to form new clusters from data points that have the close affiliation with existing classes. The clustering prototype pattern matrix is updated according to equation (6).

$$G_i^{j+1} = \left[ \frac{\sum_{j=1}^m u_{ij}^{j+1} * \sigma * f(x_j)}{\sum_{j=1}^m u_{ij}^{j+1} * \alpha^m} \right]. \quad (7)$$

Repeat the above two-step algorithm by accumulating the number of iteration steps  $j$  to  $j+1$  until the minimum value of the approximation objective function meets the requirements of equation (8), resulting in the cluster partition matrix  $G$  and the cluster statistical probability  $P$ .

$$|P^j - P^{j+1}| < \sigma * e^j. \quad (8)$$

Compared with the conventional accounting statements, it can be found that fraudulent accounting statements have specific characteristics in terms of data structure and its

correlation. The data mining technology is used to compare the data of regular and fraudulent financial statements, and the data of financial statements are processed by attribute generalization and correlation analysis technology to identify the differences of attribute characteristics between regular and fraudulent financial statements, which provides a basis for auditors to make correct judgments.

### 3. Analysis of Results

**3.1. Model Analysis.** Our result is achieved by using the FCM algorithm, in which the parameters of the modified algorithm are automatically adjusted. Balance sheet ratio, EBITDA, accounts receivable to revenue ratio, current asset turnover ratio, and cash ratio, which was used as cluster numbers and cluster indicators. In this paper, we use MATLAB software for mathematical analysis to realize the FCM algorithm to classify the accounting data of companies with accounting fraud violations and screen out the irrationalities. Figures 3(a) and 3(b) show the clustering results of two clustering targets and three clustering targets, respectively, with cluster centers 2 and 3 and C-means clustering using randomly selected seed. From Figure 3(a), it can be seen that when the C-means algorithm selects the sample data as two cluster target classification numbers for cluster partitioning, the best screening rate for identifying false financial statements is 61.05%, and the number of false financial statements in this cluster accounts for 36.21% of the total false financial statements. As shown in Figure 3(b), when the C-mean algorithm selects the sample data as three cluster target classifications for clustering, the best screening rate for identifying false financial statements is 87%, and the number of false financial statements in this cluster accounts for 32.15% of the total false financial statements. From setting the number of cluster centers, it is clear that increasing the number of cluster centers is beneficial to improve the screening rate of false financial statements.

Figure 4 shows the table of statistical characteristics of indicators obtained from the C-mean algorithm when three cluster target classifications are selected for cluster



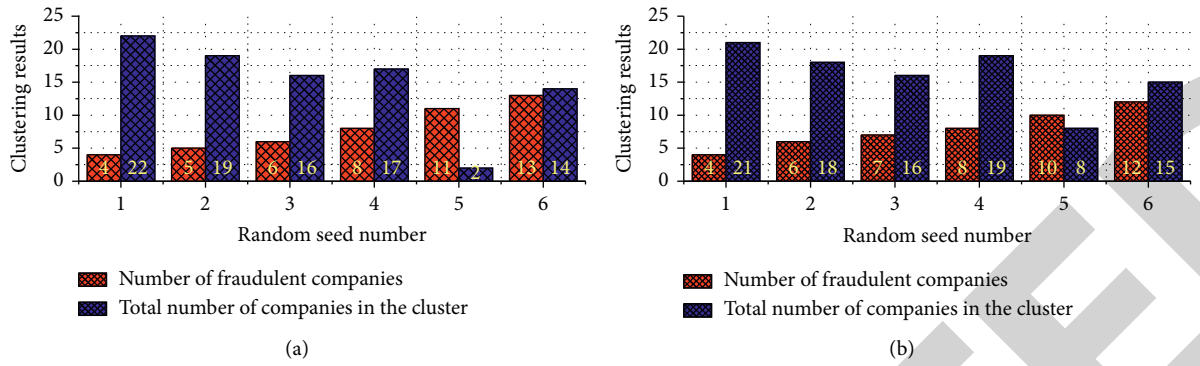


FIGURE 3: Cluster classification results. (a) Two cluster target classification numbers. (b) Three cluster target classification numbers.

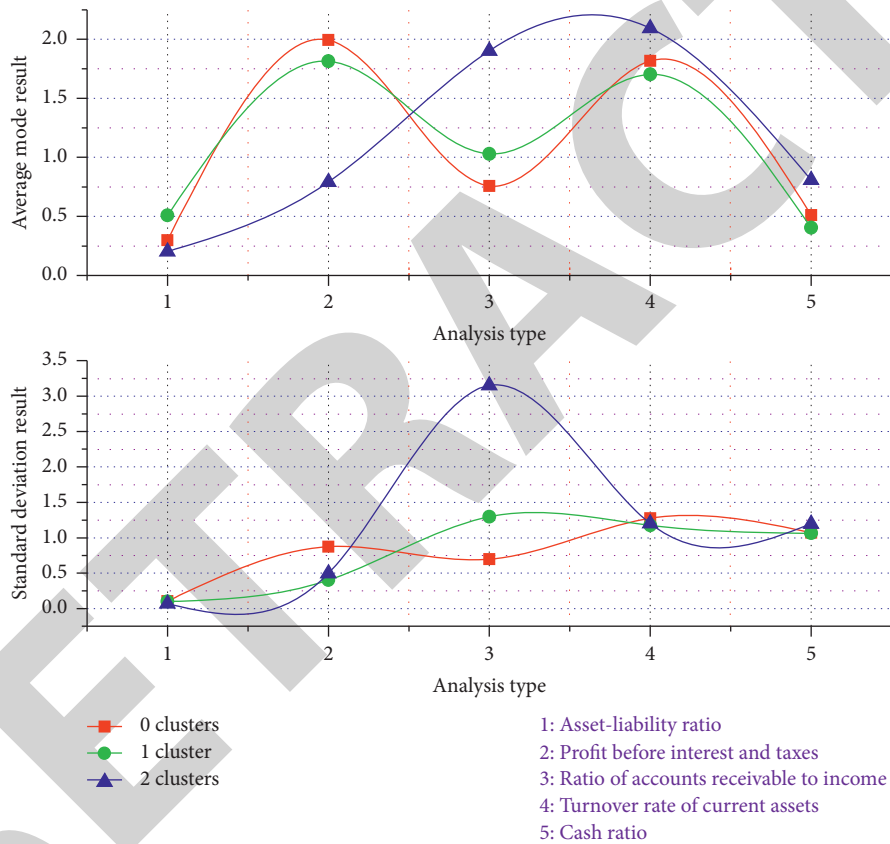


FIGURE 4: Statistical characteristics of indicators.

partitioning. From Figure 4, it can be seen that, in cluster 2, the best screening rate for identifying false financial statements is 87.51%, and the number of false financial statements in this cluster accounts for only 32.15% of the total number of false financial statements. The reason for this is that in this statistical table, the characteristics of this cluster show that the liquidity turnover ratio is higher than the average of the other two clusters, while the anomaly is that its gearing ratio and EBITDA and accounts receivable to revenue ratios are lower than the other two categories, which indicates that when a fraudulent company is in a financial crisis situation, it is likely to provide false financial reports and thus conceal the financial crisis situation.

The results of the six experiments selected for the above clustering analysis were compared and analyzed, and the results are shown in Figure 5. The clustering algorithm designed in this paper is effective, and it is feasible to use the clustering algorithm to analyze the false financial statements when the samples lack a priori knowledge or the credibility of the training samples is in doubt, which can provide help to the decision-makers, among which the samples judged as false sample class by this algorithm are also of concern and need further manual financial testing to discern whether they are falsified.

The study of the predicted sample using the established identification model was able to identify 80% of the falsified

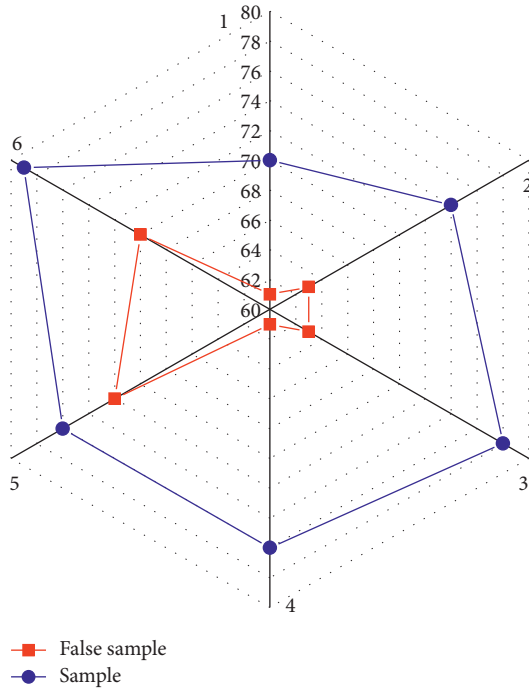


FIGURE 5: Average division accuracy of FCM algorithm experiments.

financial reports with a relatively high degree of discrimination. The continuous use of this model to analyze the latest listed companies punished by the China Securities Regulatory Commission or the Ministry of Finance shows that, in the different financial samples provided by the three types of falsified listed companies, seven different models can be judged with the help of this model, so the recognition rate reaches 87.5% and the overall discrimination is good. However, this model also has disadvantages, especially it is not able to identify financial reports with falsification problems. Therefore, the above model should be adopted with caution.

**3.2. Algorithm Effect Analyses.** In order to verify that the features selected in this paper have a better effect than the gray value features alone in identifying false financial statements, this paper uses the FCM algorithm to perform false recognition on the sample data based on these two features and examines the recognition effect under these two features, and the comparison results are shown in Figure 6.

In this paper, the threshold value is selected as 0.5, when the model output is greater than 0.5, the company is considered as fraudulent; when the model output is less than or equal to 0.5, the company is considered as nonfraudulent. As mentioned earlier, it is meaningless to judge the goodness or badness of the recognition model by the correct recognition rate of the training samples, but the recognition results of the test set samples that the model has not learned should be used as the reference for the effect evaluation. We calculate the performance of the comprehensive financial fraud recognition model in the test set and the performance of each single financial fraud recognition model in the test set

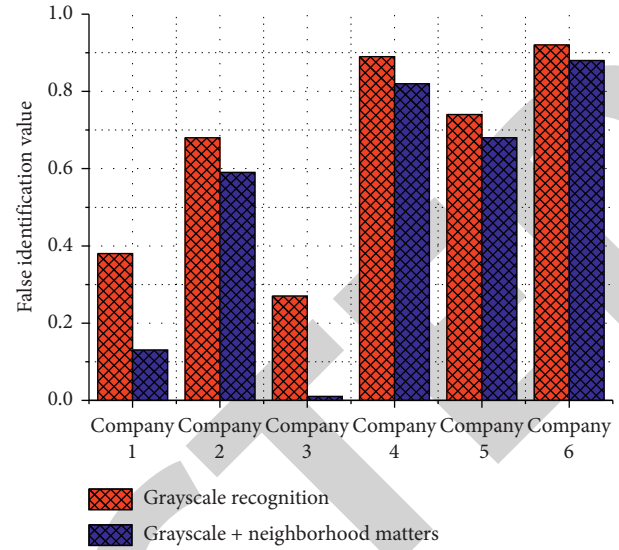


FIGURE 6: Comparison of false recognition effect under two kinds of features.

based on the statistical way of the confusion matrix and the formula of each evaluation index. The results of the calculation of the four metrics of accuracy, precision, recall, and  $F1$  score are shown in Figure 7. The results show that the integrated identification model is at a high level when compared with the single identification model for each metric. From the results of the single recognition model, it can be seen that the support vector machine-based model has the best performance in the metrics of model accuracy, recall, and  $F1$  score, reaching 74.74%, 77.51%, and 78.68%, respectively. In terms of accuracy, the BP neural network-based model performs the best with 82.13%. In comparison with the integrated model, the integrated model outperformed the highest vector machine-based model by 6.5% in the accuracy index, 75.66%, and by 1.6% in the precision index, 73.49%, and then the highest BP neural network-based model; in the recall and  $F1$  scores, the integrated model outperformed the support vector sister-based model by 7.2% and 5.8%, 69.12% and 78.68%, respectively. In terms of recall and  $F1$  score, the integrated model is 7.2% and 5% higher than the support vector model, 71.26% and 77.51%, respectively.

**3.3. Analysis of Recognition Results.** The support vector machine-based recognition model:  $AUC = 0.735$  is the best  $AUC$  value that can be derived from this model and its corresponding threshold is also the best division (Figure 8(a)). The sensitivity and specificity of the support vector machine-based recognition model are 0.784 and 0.686, respectively; the sensitivity of the BP neural network-based recognition model is 0.794, the specificity is 0.657, and the  $AUC$  area is 0.725 (Figure 8(b)); the sensitivity of the random forest-based recognition model is 0.667, and the specificity is 0.716 with an  $AUC$  area of 0.691 (Figure 8(c)). Among these three models, the  $AUC$  area of the support vector machine-based financial fraud identification model is



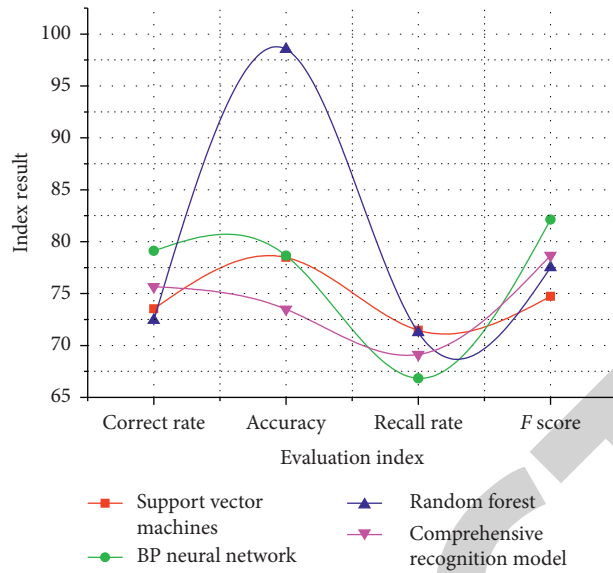
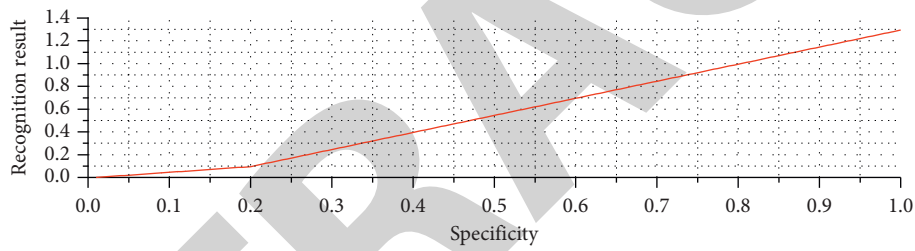
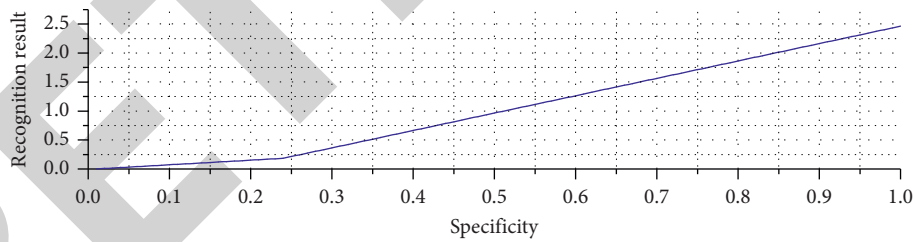


FIGURE 7: Calculated results of each model metric.



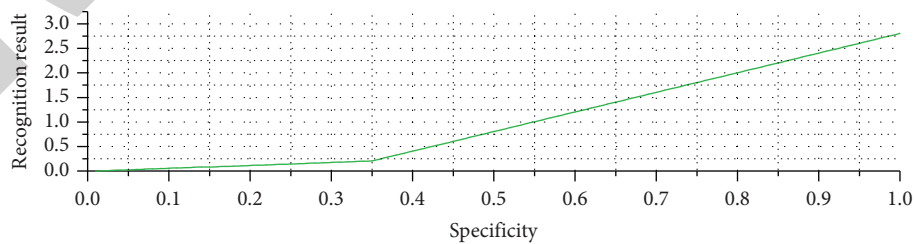
Support vector machine model ROC curve

(a)



BP neural network model ROC curve

(b)



Random forest model ROC curve

(c)

FIGURE 8: Receiver operating characteristic (ROC) curve.

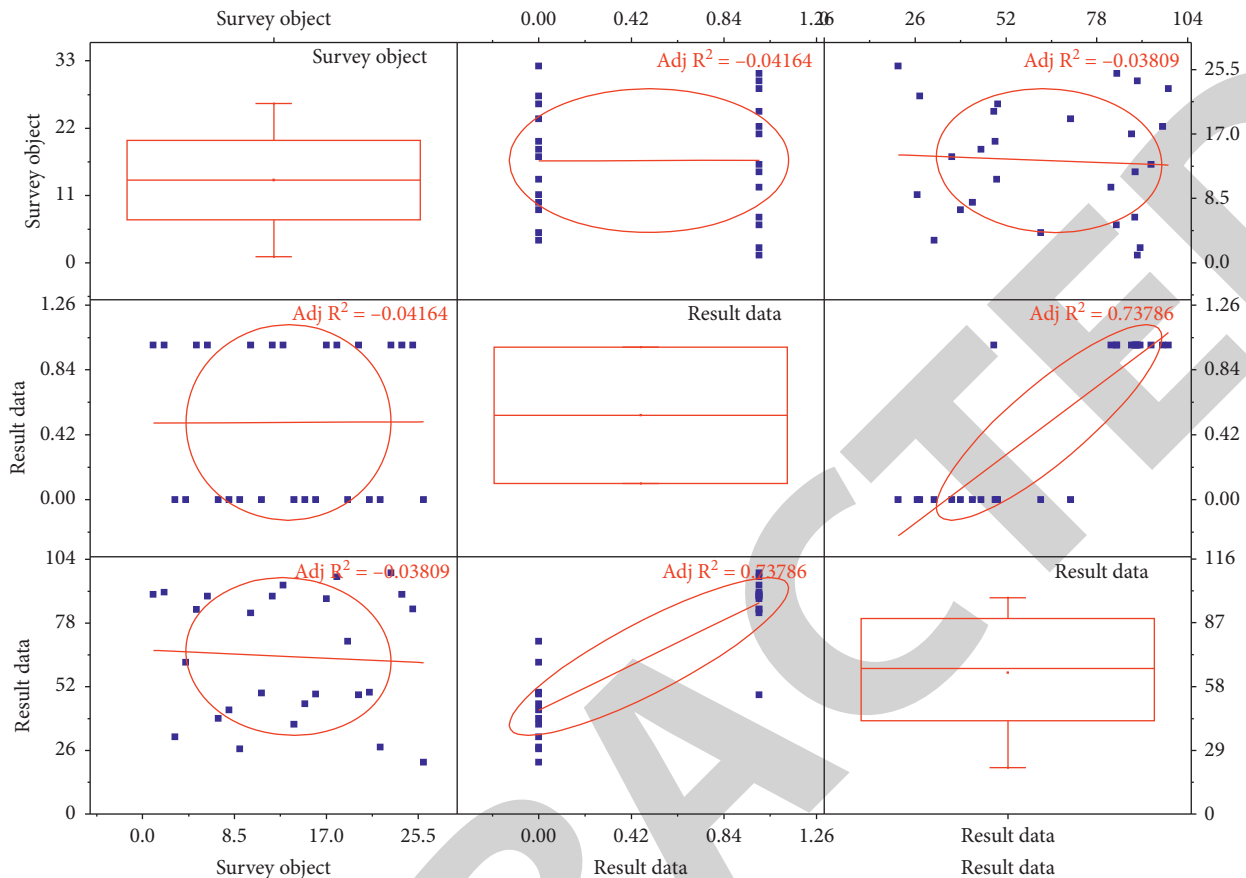


FIGURE 9: Fraud likelihood prediction.

the largest, followed by the BP neural network-based and random forest-based financial fraud identification models, respectively. In other words, among the three data mining financial fraud single identification models, the support vector machine-based financial fraud identification model performs the best, and the BP neural network-based and random forest-based financial fraud identification models rank the second and third, respectively.

After comparing the recognition effect of each single model, we cannot directly conclude which model is the best one for financial fraud recognition, because each model has its own strengths and weaknesses in addition to various conditions in practical applications. In the process of building three financial fraud identification models based on support vector machine, BP neural network, and random forest, we found that the advantage of the support vector machine-based model is that the classification method is simple and performs well with a small amount of sample data, but the recognition effect will become worse when the sample size is too large; the advantage of the BP neural network-based model is that it has good self-adaptive ability and can learn independently. The advantage of BP neural network-based model is that it has good self-adaptive ability, can learn independently, and obtains the final output result by repeatedly adjusting the error, which has good fault tolerance, but its disadvantage is that the process of model construction is more complicated, and the phenomenon of

overfitting easily occurs in the training process. The advantage of the random forest-based model is that the model is easy to understand, suitable for processing data with a large sample size, and less prone to overfitting and can determine the importance of each variable feature, but the disadvantage is that the fine adjustment of the training set data will affect the stability of the model. After considering the model recognition effect and the advantages and disadvantages of the model, this paper tries to combine three data mining techniques to form a comprehensive model for financial fraud recognition in order to achieve a better recognition effect.

The best performance of the model is achieved by integrating the BP neural network with self-encoder extracted features and the support vector machine with the significance test extracted features. The output of the logistic regression model as a secondary classifier can not only identify whether a listed company is fraudulent or not but also predict the probability of fraud, as shown in Figure 9. This will help listed company regulators and auditors to use the model prediction results for further judgment and decision-making.

In this study, the tendency is to construct a model that can predict the likelihood of corporate fraud, and the percentage results of the logistic regression output can be used to indicate the probability of corporate fraud. This indicates that although the existence of fraud is not disclosed at

present, it does not mean that there is no risk of fraud, and the supervision of companies with high possibility of fraud should be strengthened to prevent fraud before it happens, which is the significance of this study.

#### 4. Conclusion

This paper uses the FCM algorithm to identify fraudulent financial statements, analyzes the data structure of fraudulent financial statements and their specific relevance characteristics, compares the data of conventional financial statements and fraudulent financial statements, summarizes and analyzes the relevance of their data, and based on the FCM clustering method amplifies the difference of attribute characteristics, thereby identifying fraudulent financial statements and providing a basis for auditors to make correct judgments. It analyzes the data body through attribute generalization and correlation analysis. Based on HCM, FCM, and PCM clustering algorithms, a general algorithm based on data typicality, spatial structure information, and likelihood clustering information and its related objective function problems are studied, and related issues are discussed. The general algorithm combines the advantages of fuzzy mathematics to deal with the fuzziness of the data, the objectivity of the spatial neighborhood of the data affected by the neighborhood's data, and the typicality of the distance between the data and the center. Based on the means of financial fraud, this paper identifies financial fraud from two aspects, fraud characteristic signals and fraud risk factors, and proposes five targeted preventive measures based on the problems found.

#### Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

#### Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### References

- [1] S. K. Majhi, S. Bhattacharya, R. Pradhan, and S. Biswal, "Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 3, pp. 2333–2344, 2019.
- [2] A. AlShibli and H. Mathkour, "Fuzzy methods for the detection of copy number variations in comparative genomic hybridization arrays," *Saudi Journal of Biological Sciences*, vol. 27, no. 12, pp. 3647–3654, 2020.
- [3] F. Samadi, G. Akbarizadeh, and H. Kaabi, "Change detection in SAR images using deep belief network: a new training approach based on morphological images," *IET Image Processing*, vol. 13, no. 12, pp. 2255–2264, 2019.
- [4] S. K. Saeed, "A fraud-detection fuzzy logic based system for the Sudanese financial sector," *Journal of Engineering and Computer Science (JECS)*, vol. 20, no. 1, pp. 17–30, 2019.
- [5] N. Chai, B. Wu, W. Yang, and B. Shi, "A multicriteria approach for modeling small enterprise credit rating: evidence from China," *Emerging Markets Finance and Trade*, vol. 55, no. 11, pp. 2523–2543, 2019.
- [6] N. Kaja, A. Shaout, and D. Ma, "An intelligent intrusion detection system," *Applied Intelligence*, vol. 49, no. 9, pp. 3235–3247, 2019.
- [7] R. Sheeba, G. Jiji, and T. A. B. Raj, "WFCM based big sensor data error detection and correction in wireless sensor network," *Cluster Computing*, vol. 22, no. 2, pp. 3191–3200, 2019.
- [8] S. K. Saeed and H. Hagra, "A big bang–big crunch type-2 fuzzy logic based system for fraud-detection in the Sudanese financial sector," *Journal of Engineering and Computer Science (JECS)*, vol. 21, no. 3, pp. 16–28, 2020.
- [9] L. K. Singh, H. Garg, N. A. Pooja, and M. Khanna, "Performance analysis of machine learning techniques for glaucoma detection based on textural and intensity features," *International Journal of Innovative Computing and Applications*, vol. 11, no. 4, pp. 216–230, 2020.
- [10] M. Nawaz and H. Yan, "Saliency detection via multiple-morphological and superpixel based fast fuzzy c-mean clustering network," *Expert Systems with Applications*, vol. 161, Article ID 113654, 2020.
- [11] A. Arshad, S. Riaz, and L. Jiao, "Semi-supervised deep fuzzy c-mean clustering for imbalanced multi-class classification," *IEEE Access*, vol. 7, pp. 28100–28112, 2019.
- [12] C. Yuan, X. Sun, and Q. M. J. Wu, "Difference co-occurrence matrix using BP neural network for fingerprint liveness detection," *Soft Computing*, vol. 23, no. 13, pp. 5157–5169, 2019.
- [13] Y. Zhang, X. Qian, J. Wang, and M. Gendee, "Fuzzy rule-based classification system using multi-population quantum evolutionary algorithm with contradictory rule reconstruction," *Applied Intelligence*, vol. 49, no. 11, pp. 4007–4021, 2019.
- [14] Y. Zhao, J. Xu, and J. Wu, "A new method for bad data identification of oilfield system based on enhanced gravitational search-fuzzy C-means algorithm," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 11, pp. 5963–5970, 2019.
- [15] C. L. Chowdhary and D. P. Acharjya, "A hybrid scheme for breast cancer detection using intuitionistic fuzzy rough set technique," *International Journal of Healthcare Information Systems and Informatics*, vol. 11, no. 2, pp. 38–61, 2016.
- [16] M. Hao, M. Tan, and H. Zhang, "A change detection framework by fusing threshold and clustering methods for optical medium resolution remote sensing images," *European Journal of Remote Sensing*, vol. 52, no. 1, pp. 96–106, 2019.
- [17] M. Jing and Y. Tang, "A new base basic probability assignment approach for conflict data fusion in the evidence theory," *Applied Intelligence*, vol. 51, no. 2, pp. 1056–1068, 2021.
- [18] L. Deng, D. Li, X. Yao et al., "Mobile network intrusion detection for IoT system based on transfer learning algorithm," *Cluster Computing*, vol. 22, no. 4, pp. 9889–9904, 2019.
- [19] W. Jiang, C. Xie, M. Zhuang, and Y. Tang, "Failure mode and effects analysis based on a novel fuzzy evidential method," *Applied Soft Computing*, vol. 57, pp. 672–683, 2017.
- [20] R. Jing, S. Liu, Z. Gong et al., "Object-based change detection for VHR remote sensing images based on a Trisiamese-LSTM," *International Journal of Remote Sensing*, vol. 41, no. 16, pp. 6209–6231, 2020.
- [21] S. Mukherjee, I. Cheng, S. Miller, T. Guo, V. Chau, and A. Basu, "A fast segmentation-free fully automated approach to white matter injury detection in preterm infants," *Medical & Biological Engineering & Computing*, vol. 57, no. 1, pp. 71–87, 2019.

- [22] K. Benmouiza and A. Cheknane, "Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting," *Theoretical and Applied Climatology*, vol. 137, no. 1-2, pp. 31-43, 2019.
- [23] M. Ghobaei-Arani and A. Shahidinejad, "An efficient resource provisioning approach for analyzing cloud workloads: a metaheuristic-based clustering approach," *The Journal of Supercomputing*, vol. 77, no. 1, pp. 711-750, 2021.

RETRACTED