International Journal of Database Theory and Application
Vol.8, No.1 (2015), pp.157-168
http://dx.doi.org/10.14257/ijdta.2015.8.1.17

Detection of Financial Reporting Fraud Based on Clustering Algorithm of Automatic Gained Parameter K Value

Ran Li

School of Management, Hefei University of Technology, Hefei, China
ranran19780212@126.com

Abstract

Listed companies' financial reporting fraud has been a major problem in the research history of accounting. It produces an extremely bad and wide range of influence on the development of securities market. With the continuous development and progress of the stock market, the requirements for strictly controlling and preventing financial reporting fraud are also increasingly high. There have been a lot of studies of financial fraud at home and abroad. They are usually about motivations, means, identification and controlling of financial fraud. Financial fraud recognition is usually divided into signal judgment and model identification. However, the existing recognition models' accuracy is generally not high. There is a large room for improvement and the models are not applicable enough. In addition, in the era of knowledge economy, with the continuous development of information networks, computer network technology is more and more generally applied in the field of finance. Especially the use of computers in financial reporting fraud investigation can greatly reduce the manpower and resources as well as improve the efficiency of identification. But it is not known that what kind of method combined with computer technology can better identify financial reporting fraud. In this case, the paper aims at establishing an accurate financial reporting fraud recognition model based clustering method.

Keywords: *Financial reporting fraud, Clustering algorithm, Recognition model*

1. Introduction

The problem of financial reporting fraud is deep-rooted in the research history of accounting. Continuous development of the market and improvement of the system did not clearly "purge the rottenness". Continuously strengthening the supervision of financial fraud did not completely prevent financial reporting fraud. In the Early international market, financial fraud case of Enron and HealthSouth in America took place because of their business failure and also the CPA audit failure. In the domestic market, there were Yuanye management and Hongguang Industrial Management. Although they were not as notorious as Enron, they occurred much earlier and the adverse effects caused inherently are not small. Throughout the development of domestic and international securities markets, financial fraud cases can be described as endless. The occurrence of these cases caused an adverse impact which should not be underestimated in every way. Financial fraud affects the country's macroeconomic policy-making and only accurate, truthful accounting information can improve government departments' macro-control. In addition, financial fraud has a seriously negative effect on the social values and professional ethics. Financial fraud will also prejudice the interests of accounting information users, make the public doubt the integrity of accounting, shake the foundation of the credit base of the economic market and endanger macroeconomic normal operation. Therefore, study on financial reporting fraud recognition methods, accurately identify financial fraud and combat corrupt practices, lay a good foundation for the development of the market is an urgent problem to be solved.

Financial fraud incidents occurring one after another had a bad influence in the market development. Therefore, more research scholars have conducted research on the issue of financial reporting fraud. The researches mostly focused on financial reporting fraud drivers, means, identification and prevention. For the respect of motives of financial fraud, in addition to the several well-known basic theory: the iceberg theory, triangular theory, the four-factor theory and the theory of risk factors, many researchers made deeper and more detailed analysis on this basis. Deng Hua [1-2] thought that ownership structure, non-independent directors and audit workers taking part in fraud cases are several factors for financial fraud after analyzing the related theories. Yu Huiqin[3] proposed that motives of corporate financial fraud contain cutting down political cost, price manipulation, getting listed qualifications, avoiding delisting and additional allotment refinancing. Liu He [4] put forward that the drivers can be summed up into two points: improper incentives induce the self-interested behavior of executives; weak oversight mechanisms allow fraud to take advantage. Hou Xinxia [5] held the point that profit-driven is the root cause of financial fraud. In addition, there are management chaos and inadequate supervision.

From the perspective of fraud methods, Xu Xiangfeng [6] divided the financial reporting fraud into two categories: fabricate financial report; partly adjust the revenue recognition methods and change the depreciation method to distort some relevant data on the report. Ming Hongsheng [7] summarized the general expression of financial reporting fraud as following: use improper accounting policies and accounting estimations; separation, simulation and some other "accounting innovation"; fictitious economy business; asset restructuring and related transaction; tax fraud. Yue Dianmin [8] proposed that accounting fraud generally contains illegal disclosure of accounting information and financial reporting fraud. financial reporting fraud is usually brought out by fabricating net assets or profits. Chen Huixuan and Zhu Jun [9] divided financial reporting fraud into six types in the study of China's listed companies' financial reporting fraud features. They are fabricating profits, fabricating assets, delaying disclosure, misrepresentation, material omissions and irregular warranty.

This paper focuses on the establishment of recognition model of financial reporting fraud. In this respect, Christopher [10] fully took advantage of the description of triangle theory in SAS No.99 and built a more effective recognition model. Chen Pengpeng [11] analyzed the recognition method of listed companies’ financial fraud by collecting data and applying SPSS software. But the recognition model he created has a low average accuracy rate of only 70.6%. Wu Yixin [12] dealt the eligible data with logistic regression method and established a financial fraud criminal quantitative identification mode. Qiao Hong [13] identified false financial reports of listed companies by the use of GMDH model and got five key indicators which can display signs of financial fraud. The empirical research proved that GMDH model has high recognition and forecasting capability. Based on the principle of clustering, Chen Qingjie [14] used RBF neural network model and added the indicator of managers’ characteristics in the traditional detection model to improve the model and enhance its recognition ability. Ma Jingjing [15] proposed a quantitative identification method based on data mining and cluster analysis; as well as a qualitative identification method based on the vulnerability of laws and regulations, credit rating of companies. The combination of the two methods can improve financial fraud recognition ability. Kan Baokui [16] proposed a new support vector machine method which enhanced by spectral clustering. The improved model’s identification accuracy and generalization capacity is significantly better than the ordinary SVM and BP neural network. Liu Yuan [17] used clustering technique to raise the overall recognition accuracy of the model from 71.58% to 81.36%.

In today's accounting information age, companies generally manage financial information by accounting information systems and the means of fraud are more and more sophisticated. It is obviously unrealistic to complete the establishment of the model and recognition of financial fraud only by the human. So it is necessary to use computer network technology. During our research, we found that predecessors got good performance when they combined the mathematical methods with computer network technology in the establishment of financial fraud recognition model. However, the existing models are unable to meet the needs of market developing and its ability of prediction, identification needs to be improved. This paper will select appropriate indicators to establish recognition model of financial reporting fraud based clustering method. Part two is the basic idea of cluster arithmetic. Part three introduces K-means clustering algorithm and the improved K-means clustering algorithm. Simulation experiments and results analysis will be given in part 4 and we will also contrast the improved K-means clustering algorithm with the simple one by experiments. The last part is conclusions.

2. Basic Idea of Cluster Arithmetic

2.1. The Definition of Cluster

Cluster analysis means dividing a data object into multiple categories or clusters according to the principles as follows: make the data objects of the same clusters or types have a high degree of similarity; make the data objects of different clusters or types as different as possible. The mathematical definition of cluster analysis is as follows:

The studied sample set is E . C is defined as a non-empty set. $C \subset E$ and $C \neq \varphi$ .Clustering is a collection of classes $C_1, C_2, C_3, \ldots, C_k$ which can meet the following two conditions:

(1) $C_1 \cup C_2 \cup C_3 \cup \ldots \cup C_k = E$

(2) $C_i \cap C_j = \varphi \ (i \neq j)$

As it can be seen from the definition of cluster, each sample of those which are concentrated must belong to no more than one class.

The basic steps of clustering analysis are shown as Figure1: First of all, extract and select some cluster-related characteristics of the data and store them in the vector. We should also minimize the amount of information redundancy. Similarity magnanimity is used to measure the similarity between two characteristic vectors and it is generally calculated by Euclidean distance. Next is the clustering algorithm design and selection of clustering based on user requirements, different customer requirements, also with the use of clustering algorithms are not the same. Next is the design and choice based on certain requirements.

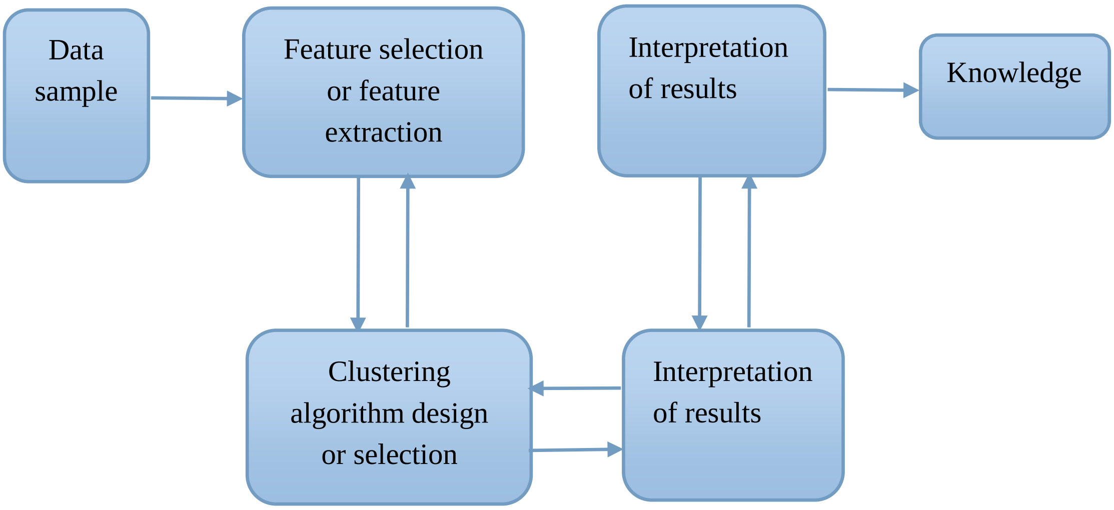
Data sample
Feature selection or feature extraction
Interpretation of results
Knowledge
Clustering algorithm design or selection
Interpretation of results

Figure 1. Steps of Clustering Analysis

2.2 Similarity Metric Function

Cluster analysis is divided according to the degree of difference between the objects. And this degree is usually measured by the "distance" which is defined as follows: assume that there are n multiple observed data sets and each of them has P properties:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T, i = 1, 2, \dots, n \text{ 。}$$

Assume that $d(x_i, x_j)$ is the distance between sample x_i and x_j, they should meet the three conditions below:

① $d(x_i, x_j) \geq 0$ and $d(x_i, x_j) = 0$ are only under the condition $x_i = x_j$;

② $d(x_i, x_j) = d(x_j, x_i)$;

③ $d(x_i, x_i) \leq d(x_i, x_k) + d(x_k, x_j)$

Users can define the "distance" according to the actual situation in the case of satisfying the above three conditions. Commonly used distances are:

① Euclidean distance

$$d_{ij}(2) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2} \tag{1}$$

② Block distance

$$d_{ij}(1) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{2}$$

③ Chebyshev distance

$$d_{ij}(\infty) = \max_{1 \leq k \leq p} \left| x_{ik} - x_{jk} \right| \tag{3}$$

④ Markov distance

$$d_{ij} = d(x_i, x_j) = \left[(x_i - x_j)^T S^{-1} (x_i - x_j) \right]^{\frac{1}{2}} \tag{4}$$

⑤ Variance weighted distance

$$d_{ij} = d(x_i, x_j) = \left[\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{s_k^2} \right]^{1/2} \tag{5}$$

2.3 Criterion Function

Cluster analysis process is actually process of pursuing the smallest convergence of the objective function. It is another key issue of cluster analysis that how to design and define such an objective function. The objective function also referred to criterion function. Selection of criteria function directly affects the quality of clustering results: if the selection is appropriate, clustering quality will be raised. Commonly used criteria functions are as following:

(1) Criterion of error square

Squared error criterion function can be used when the quantities of various types of samples don't differ significantly and the samples are intensive. In this case, it can help us get a better clustering result. Squared error criterion function is defined as follows:

$$J_c = \sum_{i=1}^{c} \sum_{k=1}^{m_i} \left\| x_k - m_i \right\|^2 \tag{6}$$

m_i is the average value of samples entity of type w_i. Its calculation method is shown below:

$$m_i = \frac{1}{n} \sum_{i=1}^{n_i} x_i,\ i = 1, 2, \dots, c \tag{7}$$

Among c sample sets, m_i respectively presents the center of i th sample set. It can also be used to present i th cluster.

(2) Criterion of weighted average squared sum of distance

$$J_i = \sum_{i=1}^{c} P S^*_i \tag{8}$$

S^*_i represents the average squared distance between samples within the class. It is calculated as follows:

$$S^*_i = \frac{n_i(n_i - 1)}{2} \sum_{x \in x_i} \sum_{x' \in x_i} \left\| x - x' \right\|^2 \tag{9}$$

n_i is the number of samples x_i. We have totally $\frac{n_i(n_i - 1)}{2}$ kinds of pair combinations. $\sum_{x \in x_i} \sum_{x' \in x_i} \left\| x - x' \right\|^2$ is the sum of clusters between samples.

3. K-means Clustering Algorithm

The purpose of K-means clustering algorithm is to divide the set $X = \{x_1, x_2, \dots, x_n\}$ which has n data objects into k classes $C_j\ (j = 1, 2, \dots, k)$. First of all, randomly select initial cluster centers of k classes. Then divide each data object in the collection into the nearest cluster center belongs to the classes and thus we have k initial cluster distributions. After the initial division of the classes, recalculate each center of them in accordance with certain rules (generally use distance).

If the calculated center is different with the former, then assign the data again and so forth iteration continues until each center of the class does not change any more (that means all the data objects have been correctly classified). Now criterion function is convergent and the algorithm terminates.

3.1. Calculation Steps of K-means Algorithm

(1) Select k initial cluster centers $C_1, C_2, C_3, \dots, C_k$ from the initial data set X as a reference randomly.

(2) Consider C_1, C_2 of $C_1, C_2, C_3, \dots, C_k$ as initial reference points, divide X according to the principles as following: if we have $d_{ie}(x_i, c_e) < d_{if}(x_j, c_f),\ j = (1, 2, \dots, k), e \neq f, i = (1, 2, \dots, k)$, then divide x_i into c_e. Else x_i will

be divided into c_f.

(3) Recalculate the center of each cluster subclass $C_1^*, C_2^*, C_3^*, ..., C_k^*$ according to the formula: $c_i = \frac{1}{n_i}\sum_{x \in w_i} x$.

(4) If any $i = (1, 2, ..., k), c_i$ can be possible, the algorithm ends up and the current c_i represents the form of the final cluster divided. Else, turn back to step2. In order not to meet the end conditions of step 4 and make it trapped in an infinite loop, we will usually pre-set a maximum number of iterations in the algorithm as the threshold.

(5) Output the final clustering result. The progresses of K-means algorithm are shown as Figure 2:

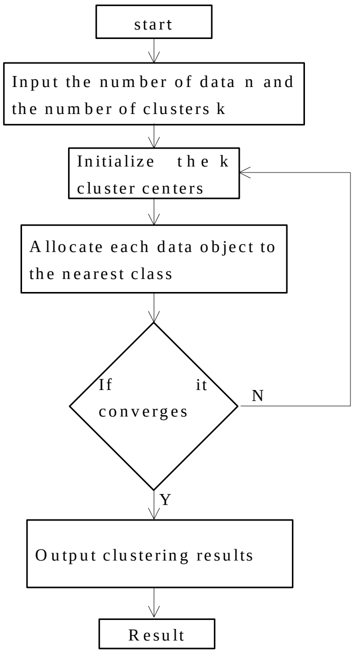
start
Input the number of data n and the number of clusters k
Initialize the k cluster centers
Allocate each data object to the nearest class
If it converges
N
Y
Output clustering results
Result

Figure 2. Processes of Clustering Algorithm

The source code of the algorithm is shown as below:

Table 3.1. Algorithms Table

```
Public void UpdateCluster()    {
        int flag=0;
        try{ {
              while(true){
                      getCluster();
                      bakCluster.clequeue();
                      for(int i=0;i<k,i++){
                                    bakCluster.addqueue(kmeansQue.getqueue(messageNo+i));    }
                      for(int i=0;i<k,i++){
                                    kmeansQue.delqueue(messageNo); }
                      flag=1;
                      for(int i=1; i<k,i++){
                      float[] x=countCenter(i);
                      B eau a=new Beau(x[0], x[1], x[2], x[3]);
                      B eau b=bakCluster.getqueue(i);
```

```
                kmeansQue.addqueue(a);
                if(!a.euqls(b)){
                    flag=0; }    }
            // flag
                if (flag==1){
                    break;}
                else{
                    for(int i=0;i<k;i++){
                        Cluster[i].clear();
                        mesCluster[i].clequeue();}      }      }
    }catch(Exception e){
     e.printStackTrace();         }    }
```

3.2 Clustering Algorithm of Automatic Gained Parameter Value k based on Maximized Distance

The idea of improved K-means algorithm automatically generating the value K is: first select the data objects as far as possible away from each other as the initial cluster centers to be divided; then according to the Euclidean distance, look for the data object furthest from the cluster center in each category; choose the data object which has maximum distance as the new cluster center and re-divide them; repeat the progress until the algorithm ends when meeting certain conditions. The number of clusters generates automatically in this process.

The progress of the algorithm:

(1) Set a collection containing n data objects s_n, $s_n = \{x_1, x_2, ..., x_n\}$. First select two data objects w, v which have furthest distance from each other as the initial cluster centers. $d_{wv} = \max\{d_{ij}, i, j \in 1, 2, ..., n\}$, and set $x_1^* = x_w, x_2^* = x_v, d_{wv} = d_1^*$;

(2) Calculate according to Euclidean distance and divide other $n-2$ data objects in s_n around x_1^*, x_2^*. That is $\forall i \in \{1, 2, ..., n / w, v\}$. If $\left|x_i - x_1^*\right| < \left|x_i - x_2^*\right|$, then divide x_i into x_1^*, or else divide it into x_2^*. Now we have divided s_n into two parts respectively around x_1^* and x_2^*. Record them as s_{21}^*, s_{22}^*.

(3) Calculate the distance between the data objects in s_{21}^* and x_1^*, then get $d_{21} = \max\{\left|x_i - x_1^*\right|, x_i \in S_{21}^*\}$. Calculate the distance between the data objects in s_{22}^* and x_2^*, then get $d_{22} = \max\{\left|x_i - x_2^*\right|, x_i \in S_{22}^*\}$. Make $d_2^* = \max\{d_{21}, d_{22}\}$ and record the corresponding data object as x_3^*.

(4) If $d_2^* > hd_1^*$ (h is an input parameter and is generally produced by experience.), hold x_3^* as the third cluster center. Divide s_n into three parts respectively around x_1^*, x_2^* and x_3^*. Record them respectively as s_{31}^*, s_{32}^* and s_{33}^*.

(5) Similarly repeat. If $d_{i=2}^* > h * average(d_i^* + d_{i+1}^*)$, then hold x_4^* as the forth luster center, turn back to step4. Or else the algorithm terminates and the final cluster centers are x_1^*, x_2^*, x_3^*.

The progress of the improved K-means algorithm is shown as Figure 3:

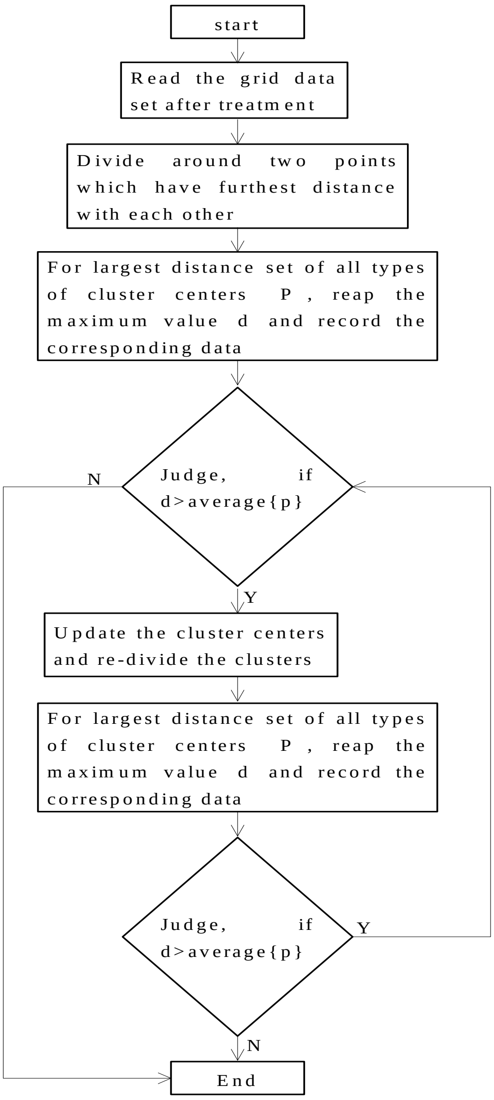
start
Read the grid data set after treatment
Divide around two points which have furthest distance with each other
For largest distance set of all types of cluster centers P, reap the maximum value d and record the corresponding data
Judge, if d>average{p}
N
Y
Update the cluster centers and re-divide the clusters
For largest distance set of all types of cluster centers P, reap the maximum value d and record the corresponding data
Judge, if d>average{p}
Y
N
End

Figure 3. Progress of the Improved K-means Algorithm

4. Simulation Experiments and Results Analysis

(1) Sample selection

Clustering is an unsupervised learning data mining and we do not need to know the data attributes in advance. But in order to examine the effectiveness of clustering results, we still need to determine if the sample is true. In accordance with the selection criteria of false samples, when the audit opinion is "declined to comment" or “adverse opinion” or “disclaimer of opinion”, we consider the report as the sample of false financial report. According to the views of the annual audit report during 2011-2012, nearly 110 companies’ financial reports meet the conditions above. We randomly select 35 false financial report samples from them. Furthermore, according to some data of 2008-2011published by China Securities Regulatory Commission on the website, we choose 20 companies which got administrative punishment by the securities and Futures Commission because of financial fraud. And the audit opinions for the annual financial report of the same year are the "standard unqualified opinion". We make the former 35 companies to be the first kind of false sample and the later 20 companies to be the second kinds of false sample.

(2)Variable selection

We choose 16 indexes related to profitability, operation and some other abilities. After processing the data, we find that there's high correlation between some indicators. So they're rejected. Finally, we have only 7 indicators which are the rate of return on net assets, assets liabilities ratio, interest coverage ratio, liquidity ratio, accounts receivable turnover rate, growth rate of net profit, capital appreciation rate. Compute the corresponding indexes of samples, use the improved K-means clustering algorithm, the experimental results obtained is shown as follows:

Table 2. Experimental Result of Test Samples

The number of experiments	1	2	3	4	Total
Predictive number of false reports in the first category	30	32	31	32	125
Accuracy of estimation	85.71%	91.43%	88.57%	91.43%	89.28%
Predictive number of false reports in the second category	14	15	16	14	59
Accuracy of estimation	70%	75%	80%	70%	73.75%
Average accuracy	80%	85.45%	85.45%	83.64%	83.64%

Table 3. Experimental Result of Control Samples

The number of experiments	1	2	3	4
Predictive number of false reports in the two categories	46	48	45	48
Accuracy of estimation	83.64%	87.27%	81.82%	87.27%

According to the experimental results of test samples, we can get that the average predictive accuracy based on the improved K-means clustering algorithm is 89.28%. In addition, in order to contrast the effect of the improved K-means clustering algorithm with simple K-means clustering algorithm, we also get the experimental result based on simple K-means clustering algorithm. As Table 4 and Table 5 show, compared to the improved K-means clustering algorithm; the average accuracy of the simple K-means clustering algorithm is much lower.

Table 4. Experimental Result of Test Samples

The number of experiments	1	2	3	4	Total
Predictive number of false reports in the first category	26	28	26	25	105
Accuracy of estimation	77.14%	80%	77.14%	71.42%	75%
Predictive number of false reports in the second category	12	13	13	11	49
Accuracy of estimation	60%	65%	65%	55%	61.25%
Average accuracy	69.09%	75.54%	70.90%	65.45%	70%

Table 5. Experimental Result of Control Samples

The number of experiments	1	2	3	4
Predictive number of false reports in the two categories	41	42	41	43
Accuracy of estimation	74.55%	76.36%	74.55%	78.18%

5. Conclusions

From the experimental results of this view, we can get some results. Firstly, the seven selected indexes have close contact with financial fraud. Secondly, the clustering model we designed is effective. It can be used divide the false financial report when people lack some necessary prior knowledge of samples or the training samples' credibility is suspected. It can provide the decision-makers some help. Thirdly, control samples may not be true, but that does not mean that our discussion about the control sample is pointless. The accuracy of the improved algorithm to detect false financial reaches 83.64%, which means that's meaningful. Finally, the accuracy of dividing the second kind of false samples in this paper is always lower than that of the first kind. It means that the deception and concealment of the second kind are both higher. We should pay much attention to those samples which are judged as false by our algorithm. The algorithm also has some limitations and it can be improved with the development of academic research and technology.

Acknowledgments

This paper belongs to the project of "the research of multi-source information service system under the cloud computing environment", supported by Chinese National Natural Science Foundation (No. 71131002) and the project of "the Empirical research of the Increase value distribution theory for Chinese Listed Companies Based on stakeholder perspective" supported by Chinese Universities Specialized Research Foundation for the Doctoral Discipline (PhD supervisor category) (No. 20100111110015).

References

[1] D. Hua and Y. Yong, "Analyses of the Factors Led to the Financial Fraud of Listed Companies in Our Country", Journal of Central South University of Forestry & Technology (Social Sciences), vol. 3, **(2013)**, pp. 92-93.

[2] L. Meng, "Analysis of Financial Fraud Motivation of listing Corporation in China", Foreign Economic Relations & Trade, vol. 7, **(2012)**, pp. 126-128.

[3] Y. Huiqin, "Research on financial fraud motivation of listing Corporation", China Business & Trade, vol. 11, **(2009)**, pp. 231-232.

[4] L. He, "Analysis of Financial Fraud Motivation", China Market, vol. 14, **(2010)**, pp. 119.

[5] H. Xinxia and H. Shuhua, "Analysis of false financial report", Chinese Agricultural Accounting, vol. 10, **(2005)**, pp. 33-34.

[6] X. Xiangfeng, "Expression and recognition method of false financial statements", Fortune World, vol. 3, **(2011)**.

[7] M. Hongsheng, "Analysis of expressions and causes financial", Pioneering with Science & Technology Monthly, statements fraud, vol. 9, **(2007)**, pp. 49-50.

[8] Y Dianmin, H. Chuanmo, W. Xiaodan and C.-H. Chu, "The Empirical Research on the Accounting Fraudulent Approaches of Public Firms in China", Auditing Research, vol. 5, **(2009)**, pp. 82-89.

[9] C. Huixuan and Z. Jun, "An Analysis of Financial Report Fraud of Listed Companies in China", Taxation and Economy, vol. 2, **(2013)**, pp. 52-57.

[10] C. J. Skousen, K. R. Smith and C. J. Wright, "Detecting and Predicting Financial Statement Fraud: The Effectiveness of the Fraud Triangle and SAS No. 99", Working Paper, Utah State University, **(2008)**.

[11] C. Pengpeng, "The Identification of the Financial Reporting Fraud of Listed Companies Based on SPSS", Value Engineering, vol. 22, **(2013)**, pp. 207-209.

[12] W. Yixin, F. Yuan and L. Shufu, "Research on determinant of Financial fraud of listing Corporation in China -- Based on logit analysis", Modern Business Trade Industry, vol. 24, **(2009)**, pp. 206-207.

[13] Q. Hong and H. Changzheng, "GMDH Identification Models of Chinese Listed Company's Financial Reports Fraud", Soft Science, vol. 1, **(2007)**, pp. 45-48.

[14] C. Qing-jie, "Research on the Improving to the Discriminating Model of the Financial Statement Fraud Based on the Characteristic of the Managers: Evidence from Chinese Listed Companies", On Economic Problems, vol. 8, **(2012)**, pp. 118-122.

[15] M. Jingjing, "Analysis and management of listing Corporation financial reporting fraud", Friends of Accounting, vol. 23, **(2012)**, pp. 95-97.

[16] K. Baokui, L. Zhixin, S. Xiaodong and Y. Zhong, "Improved Support Vector Machine Algorithm for Fraudulent Financial Statements Detection", Management Review, vol. 5, **(2012)**, pp. 144-153.

[17] L. Yuan, L. Guoqiong, W. Feng, Y. Wanli, Y. Fan, T. Xing and W. Xi, "Mining Association Rules of False Financial Statements Based on Inflection Points Interval Partition Technology", Science Mosaic, vol. 2, **(2013)**, pp. 240-244.

Author

Ran Li, he received his master's degree in Master of management, Ph. D. student. His research direction is environmental cost assessment and information management. Now he is an associate professor of Anhui Jianzhu University.