Research on Recognition of Whitewashing Degree of Financial Statements Combined with Excess-MAD Algorithm and Clustering Algorithm Model

Tingyu Luo[1]

{ m13604042455@163.com [1] }

School of management, Minzu University of China, BeiJing, China, 100081[1]

Abstract. Using the improved Excess-MAD algorithm based on Benford's law to establish financial statement whitewashing identification indicators, screen out 15% of the financial statement data available from all listed companies in the Chinese capital market from 2000 to 2021, about 14,000 pieces, and divided into three groups. Then, according to relevant financial theories, 11 financial indicators of 5 theoretical analysis dimensions are selected to establish a preliminary indicator set, and 8 indicators are reserved for modeling after being screened by Kendall correlation analysis. Finally, K-means, Gaussian Mixture Model and other clustering methods were used to establish the final binary clustering model, and Akaike information criterion and Bayesian information criterion were used for extended evaluation. The results of the model run show that the clustering algorithm model established through this process can well identify the degree of financial whitewashing of the company's statements.

Keywords: Degree of whitewashing of financial statements; Benford's law; K-means model; Excess-MAD method; Gaussian mixture model

1 Introduction

The information quality of financial statements publicly disclosed by listed companies is one of the important regulatory objects of official regulatory activities in the capital market. As an important part of the artificial manipulation of financial statements disclosed by listed companies, the whitewashing and adjustment behaviors of corporate financial personnel on the public financial statements disclosed by the company will cause distortion of the financial information disclosed by the company and reduce the accuracy of the financial statements and data quality, which in turn affects the use value of financial information and the judgments made by financial information users on this basis [1], and its overall negative impact cannot be underestimated. Therefore, compared with the financial fraud that has received more attention, the identification of the degree of whitewashing in corporate financial statements also has certain research value.

The difficulty of research on the degree of whitewashing of financial statements lies in the lack of a universal, clear and convincing criterion. However, considering that the whitewashing and adjustment of corporate financial statements is very similar to the theoretical analysis process of corporate financial statement fraud, when studying the identification of the degree of whitewashing of corporate financial statements, it should be

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334343

possible to start from the research perspective of the relationship between the numerical laws of corporate financial statements.

The rest of this paper is as follows: Section two sorts out some relevant research results, including the literature related to Benford's law and the literature related to the application of clustering algorithms in corporate financial analysis. The third section is based on previous research, establishes an evaluation index model of corporate financial statements using the Excess-MAD algorithm [2] with Benford's law as the core, and then uses it to analyze the overall data sample. A subset of samples is constructed for use by the machine learning model. In addition, using the relevant theoretical knowledge of financial management, a set of quantitative financial analysis index data sets related to the degree of corporate financial whitewashing based on sample set data was established, and the kendall correlation coefficient was used to screen them to determine the final subset of indicators for modeling. In the fourth section, descriptive statistics and cluster modeling are carried out on the data. A series of analysis and evaluation indicators are also applied, combined with relevant theoretical knowledge of finance disciplines, to evaluate, explain and analyze the actual prediction effect of the model. analyze. Two machine learning evaluation models with certain practical value in different usage scenarios were obtained. Section five summarizes the full text.

2 Sorting out related research work

2.1 Benford's law and identification of whitewashing degree of corporate financial statements

The research object of Benford's law is the distribution of the first digits in a large number of "natural" data sets generated by the carry counting system. According to this law, if the actual distribution of the first digits in a data set is inconsistent with the theoretical prediction made by the law, then It can be determined that the data set has a certain risk of being tampered with. The work of Ma Boqiang et al. (2019) has proved that Benford's law is an inherent property of all carry counting systems, and any data set based on the "natural" of the carry counting system should conform to Benford's law and its extended description[3] . This means that since the financial statement data of an enterprise is also formed by the carry counting system, such data must also conform to a series of inherent laws contained in the carry counting system itself. In practice, Benford's law has become a common method to study corporate financial fraud before the theoretical foundational loopholes are filled. Varian (1972) took the lead in proposing the idea of using Benford's law to verify the practicality and reliability of data in the field of social sciences [4]. Carslaw (1988) verified this assumption for the first time, and successfully proved the feasibility of applying Benford's law to data quality assessment in the accounting field [5]. In recent years, many Chinese scholars have studied the method of evaluating the quality of financial statements using Benford's law. For example, Ding Guoyong et al. (2003) took the lead in applying Benford's law to the audit research of financial data of Chinese universities and proved the feasibility of this method. [6]. Wan Yufei et al. (2012) used Benford's law to analyze the 2011 financial data disclosed by Chinese listed companies, and verified the effectiveness of Benford's law in finding signs of corporate financial statement manipulation in the Chinese capital market [7]. In summary, using Benford's law can successfully test whether the corporate financial statements disclosed by

Chinese listed companies are artificially manipulated, and then provide effective evaluation information related to the information quality of listed companies' financial statements.

2.2 Feasibility of the application of clustering algorithm models and other machine learning algorithms in financial information manipulation recognition models

Among the clustering algorithms, the oldest and most widely used algorithm is the k-means algorithm, which was proposed by MacQueen in 1967 [8]. The main idea of the k-means algorithm is to gradually optimize the clustering results, and redistribute the target data set to each clustering center in an iterative manner, so as to try to make the output results converge to the optimal solution.. Many scholars have applied various clustering algorithms represented by the k-means algorithm to the identification of corporate financial information manipulation. Wu Songshi et al. (2017) proposed a corporate false information identification method based on the improved k-means algorithm [9]. Guo Yang et al. (2019) used k-means clustering algorithm and hierarchical clustering algorithm to construct an evaluation system for evaluating electric power enterprises [10]. Shen Ruyuan et al. (2014) used the k-means clustering algorithm to establish an effective corporate credit evaluation and analysis system that combines corporate financial indicators [11]. Bao Xinzhong (2012) used the particle swarm-based k-means algorithm combined with rough set theory to establish a corporate financial early warning model with good prediction effect [12]. In addition, in order to solve the problem of poor performance of traditional clustering algorithms for unbalanced data sets. Cao Peng et al. (2014) proposed an improved algorithm ARSGOS based on the Gaussian mixture model, which effectively improved the recognition rate of the minority class and the overall clustering performance of the clustering algorithm model in the binary classification scenario [13]. Since the research field of corporate financial information manipulation has a great correlation with the two research fields of corporate financial early warning and corporate credit rating [14]. Therefore, the existing research results have proved the feasibility of using clustering algorithm to build a financial information manipulation recognition model based on corporate financial indicators.

3 Model variable design

3.1 Evaluation index of financial whitewashing based on Benford's law

Benford's law was first discovered by American mathematician and astronomer Simon Newcomb, and then the American physicist Frank Benford gave an empirical description from the statistics [15], and finally completed the mathematical proof by Ma Boqiang et al (2019) [3]. According to these research results, Benford's law is an objective law inherent in all base counting systems. This law describes the following phenomenon, that is, when the amount of data in a "natural" distribution data set is large enough, the distribution frequency of the first digit of the overall data is not evenly distributed, but a monotonous decreasing trend that conforms to certain mathematical laws. The inductive mathematical formula that describes Benford's law in the decimal numbering system is as follows:

$$P(d) = \log_{10}(1 + \frac{1}{d}) \ . \tag{1}$$

The value range of d is a natural number, and P(d) is the occurrence probability corresponding to the number d.

Although Benford's law has been proven to be applicable to identify traces of human manipulation of corporate financial data, the classic Benford's law also has the disadvantage of requiring a large number of data samples in its application. Generally speaking, the application threshold of Benford's law is 5,000 non-zero data, and the amount of non-zero data that a single enterprise report can provide is basically about 100, which means that it is difficult to use the classic Reliable in-depth analysis of Benford's Law. In order to solve the problem that the classic Benford law cannot properly handle small sample size data, Barney, Bradley J and Schulzke, Kurt S (2016) proposed an improved Excess MAD value evaluation method based on the classic Benford law, which can well solve Analytical problems with small data samples [2]. The mathematical formula for this method is as follows:

$$ExcessMAD \approx MAD - E(MAD) . \tag{2}$$

Among them, $E(MAD) = \sum_{k=10}^{99} \sum_{j=0}^{N} \binom{N}{j} (p_k)^j (1-p_k)^{N-j} \frac{\left|\left(\frac{j}{N}\right)-p_k\right|}{90}$, and $p_k = \log_{10}(1+\frac{1}{k})$; N is the sample size, that is, the total number of non-zero data entries; k is the data The first two digits, the value range is 10-99. In particular, when N is less than or equal to 5000, $E(MAD) \approx \frac{1}{\sqrt{158.8N}}$, therefore, when using this method in data samples less than 5000, you can directly use the following formula for calculation:

$$ExcessMAD \approx MAD - \frac{1}{\sqrt{158.8N}} \tag{3}$$

In this formula, $MAD = (\sum_{k=10}^{99} \frac{|Obs_k - Exp_k|}{N})/90$,, where Obs_k is the top The actual frequency of two digits, Exp_k is the theoretical probability of the first two digits calculated according to Benford's law. In addition, the work of Wang Jiamin et al. (2022) pointed out that the data sample size required by ExcessMAD can be reduced to a non-zero sample size range of about 100, which is the magnitude of data contained in a single annual company report [16]. The interpretation method for the calculation results is also relatively simple. If the obtained ExcessMAD value is less than or equal to 0, it means that the possibility of simple human manipulation of the data sample is very low. When it is greater than 0, it indicates that the data sample is likely to be manipulated, and the larger the value of ExcessMAD , the more serious the degree of manipulation and modification of the data sample.

3.2 Screening of financial indicator variable groups

According to the relevant financial management theory, this paper preliminarily selects 13 financial analysis indicators in five dimensions: profitability indicators, solvency indicators, operating ability indicators, business model identification indicators and cash flow indicators. Profitability indicators reflect the company's ability to obtain profits by virtue of its own assets, and are the core indicators for evaluating the company's value. Therefore, such indicators are most prone to financial whitewashing operations. Debt solvency is an important indicator for creditors and shareholders to evaluate a company's financial status, reflecting the company's ability to pay off existing debts and refinance. The management of some financially distressed companies may have an incentive to whitewash such indicators in order to conceal the real situation of the company. The operating capacity index reflects the utilization efficiency of the company's own assets, reflects the company's management level to a certain extent, and has a certain relationship with the company's solvency and profitability. To a certain extent, business model identification indicators reflect the traces of certain financial whitewashing behaviors in the company's statements. Such indicators can be used to assist in the research of whether there are certain types of financial whitewashing in the company's statements. The cash flow indicator reflects the company's cash health, and it is also the type of indicator that is relatively difficult to manipulate information through legal means. Therefore, this indicator can be used to assist in the analysis of whether other indicators are whitewashed.

The 13 indicators of the five dimensions initially screened are:

Profitability indicators: (1). Return on net assets (2). Return on total assets (3). Operating profit rate (4) Return on capital;

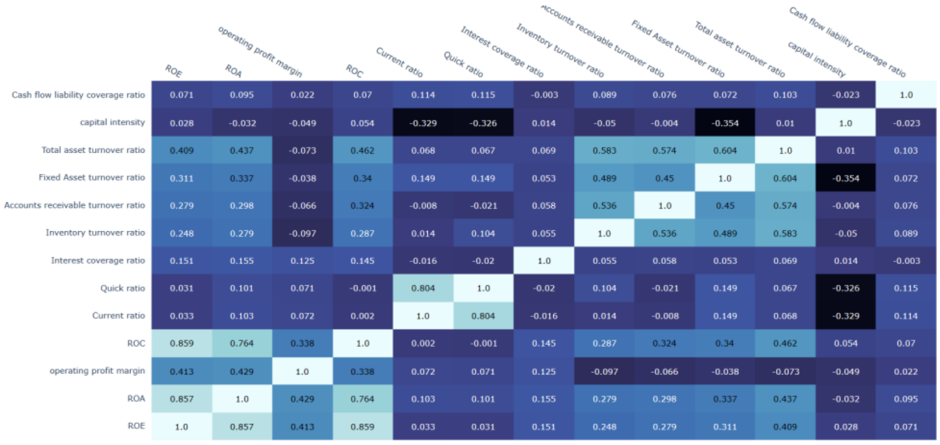
Indicators of solvency: (5) Current ratio (6). Quick ratio (7). Interest coverage ratio;

Operating capacity indicators :(8) inventory turnover rate (9). Accounts receivable turnover rate (10). Fixed asset turnover rate (11). Total asset turnover rate

Business Model Recognition Index :(12) Capital Intensity

Cash flow indicators: (13) Cash flow liability coverage ratio

Then python was used for data processing of these 13 indicators. By using the pandas.corr function, the Kendall method is selected for the calculation method parameters, and the nonlinear correlation coefficient is calculated and some indicators with too high correlation coefficients are eliminated. Finally, the values of eight retained indicators are listed in Table 1. In addition, in order to facilitate data interpretation, the calculation results of the correlation coefficient are also drawn into the correlation coefficient heat map shown in Figure 1.



	ROE	ROA	operating profit margin	ROC	Current ratio	Quick ratio	Interest coverage ratio	Inventory turnover ratio	Accounts receivable turnover ratio	Fixed Asset turnover ratio	Total asset turnover ratio	capital intensity	Cash flow liability coverage ratio
Cash flow liability coverage ratio	0.071	0.095	0.022	0.07	0.114	0.115	-0.003	0.089	0.076	0.072	0.103	-0.023	1.0
capital intensity	0.028	-0.032	-0.049	0.054	-0.329	-0.326	0.014	-0.05	-0.004	-0.354	0.01	1.0	-0.023
Total asset turnover ratio	0.409	0.437	-0.073	0.462	0.068	0.067	0.069	0.583	0.574	0.604	1.0	0.01	0.103
Fixed Asset turnover ratio	0.311	0.337	-0.038	0.34	0.149	0.149	0.053	0.489	0.45	1.0	0.604	-0.354	0.072
Accounts receivable turnover ratio	0.279	0.298	-0.066	0.324	-0.008	-0.021	0.058	0.536	1.0	0.45	0.574	-0.004	0.076
Inventory turnover ratio	0.248	0.279	-0.097	0.287	0.014	0.104	0.055	1.0	0.536	0.489	0.583	-0.05	0.089
Interest coverage ratio	0.151	0.155	0.125	0.145	-0.016	-0.02	1.0	0.055	0.058	0.053	0.069	0.014	-0.003
Quick ratio	0.031	0.101	0.071	-0.001	0.804	1.0	-0.02	0.104	-0.021	0.149	0.067	-0.326	0.115
Current ratio	0.033	0.103	0.072	0.002	1.0	0.804	-0.016	0.014	-0.008	0.149	0.068	-0.329	0.114
ROC	0.859	0.764	0.338	1.0	0.002	-0.001	0.145	0.287	0.324	0.34	0.462	0.054	0.07
operating profit margin	0.413	0.429	1.0	0.338	0.072	0.071	0.125	-0.097	-0.066	-0.038	-0.073	-0.049	0.022
ROA	0.857	1.0	0.429	0.764	0.103	0.101	0.155	0.279	0.298	0.337	0.437	-0.032	0.095
ROE	1.0	0.857	0.413	0.859	0.033	0.031	0.151	0.248	0.279	0.311	0.409	0.028	0.071

Fig. 1. Heat map of original financial indicator kendall correlation coefficient

Table 1. Definition and calculation formula of financial indicator variable group

Indicator dimension	Indicator name	Indicator calculation formula
Profitability indicator	Return on Equity (ROE)	Net Profit/Net Assets
Profitability indicator	Margin Operating	Profit/Operating Income
Solvency indicator	Current Ratio	Current Assets/Current Liabilities
Solvency indicator	Interest Coverage	Earnings Before Interest and Tax (EBIT)/Interest Expense
Operating capacity indicator	Accounts receivable turnover ratio	Operating income/net accounts receivable
Operating capacity indicator	Fixed asset turnover rate	Operating income/net value of fixed assets
Business Model Recognition	capital intensity	net value of fixed assets / Net Assets
Cash flow indicator	Cash Flow Liability Coverage Ratio	Net cash flowfrom operating activities/current liabilities

3.3 Screening and cleaning of data samples

This paper reasonably assumes that the whitewashing behavior of the company's financial statements in the current year is only directly related to the financial performance of the year. Therefore, each piece of data in the data sample consists of all the data of the balance sheet, cash flow statement, and income statement publicly disclosed by the company in the current year, as well as the financial analysis indicators and financial whitewashing indicators calculated based on the data of these three statements. The data source of this article is the original data of the financial statements of A-share and B-share listed companies in 2000-2021 available in the CSMAR Database, including the parent company's statements and consolidated statements, a total of 101,728 records. Afterwards, the necessary cleaning work was performed on the data, and the data containing zero/infinity values in the calculated financial indicators were eliminated, and the total amount of remaining data after processing was 88,840.

Then, the normality test is carried out on the data of the whitewashing index of the financial statement. The test function is the anderson function of the scipy.stats module. The calculated statistical value is about 1438, which is greater than the 1% level of the sample size. The judgment value is 1.092, which proves that the data is very typical normal distribution data. Therefore, the 5% data with the highest degree of whitewashing, the 5% data with the lowest degree of whitewashing, and the 5% data with the middle degree of whitewashing in the sample are screened out (the data with the middle degree of whitewashing means that the Excess-MAD value is between 47.5%- 52.5% of the overall score level of samples), a total of 14008 data for data modeling.

In addition, in order to improve the efficiency and accuracy of model training, the StandardScaler function of python's sklearn.preprocessing module is also used for data normalization before officially starting to train the model. The mean value of the finally obtained data distribution form is zero. The variance value of the finally obtained data distribution form is 1. Its calculation formula is

$$X_{scale} = \frac{X - X_{mean}}{S} \tag{4}$$

Among them, X is the original data, X_{mean} is the mean value of the data, S is the standard deviation of the data, and X_{scale} is the normalized data output result.

4 Model construction and related performance analysis

4.1 Raw data descriptive statistics

Firstly, the descriptive statistics shown in Table 2 are performed on the original data of the three groups of data, and the comparative radar chart shown in Figure 2 is drawn, in which the first group is the group with the lowest degree of financial whitewashing, while the third group is the group with the highest degree of financial whitewashing group. The data shows that among the average indicators, the average of the two types of indicators, the operating profit rate and the current ratio, are roughly negatively correlated with the degree of financial whitewashing, while the turnover ratio of accounts receivable is roughly positively correlated with the degree of financial whitewashing. The rest of the indicators have no obvious correlation. However, there are outliers: there are obvious abnormalities in the fixed asset turnover ratio index value of the group with the highest degree of whitewashing, and the index value of interest coverage ratio and cash flow liability ratio data of the group with the lowest degree of whitewashing. However, the average indicators of ROE and capital intensity have no obvious rules and abnormalities.

While observing the variance indicators, it can be found that except for the three indicators of return on net assets, fixed asset turnover ratio, and cash flow-to-liability ratio, the rest of the indicators follow the rule that the higher the degree of whitewashing, the greater the variance within the group. When only comparing the group with the lowest degree of whitewashing and the group with the highest degree of whitewashing, the two groups also follow the rule that the higher the degree of whitewashing, the larger the variance within the group. This result is basically in line with the prediction of the relevant analysis theoretical framework of the discipline of financial management, that is, as the degree of whitewashing of financial statements increases, the differences in the combination of whitewashing techniques of financial statements among different individuality begin to increase, and the distribution of nominal data of financial statements among individuality within a group becomes larger. This also proves that the sample data is suitable for the application of clustering algorithm for analysis and modeling.

Table 2. Descriptive statistics and comparison of the three groups of data

	Return on Equity (ROE)	Margin Operating	Current Ratio	Interest Coverage	Accounts receivable turnover ratio	Fixed asset turnover rate	capital intensity	Cash Flow Liability Coverage Ratio
Mean_1	0.206	-0.201	1.598	-1.819	259.242	84.197	0.507	2.412
Mean_2	0.261	9.631	2.740	-337.338	2042.297	79.738	0.442	1.197
Mean_3	0.258	28.556	3.210	-217.177	6414.193	6099.979	0.498	1.641
Std_1	1.280	305.622	1.223	325.272	4797.287	2636.84	1.035	37.650
Std_2	12.440	320.511	5.172	17663.973	74684.158	1073.895	2.825	19.556
Std_3	4.500	1908.74	8.8978	19935	194578	204038	3.649	52.679

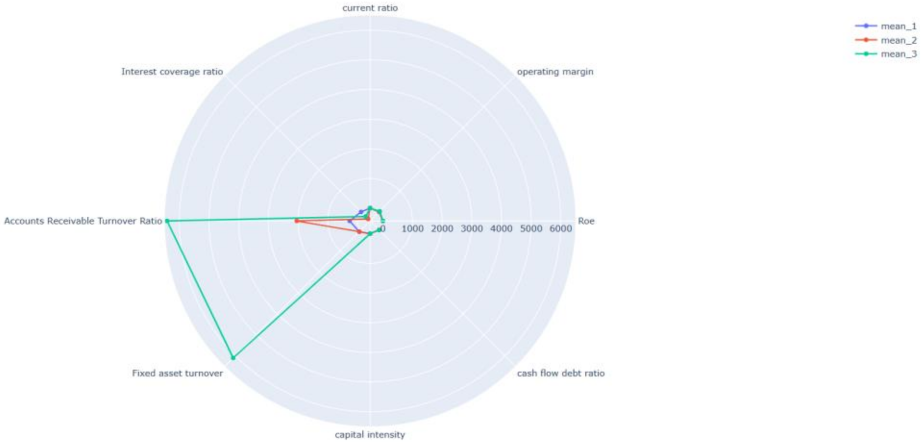


Fig. 2. Comparison radar chart of the mean values of various indicators of the three

4.2 Clustering Algorithm Model

This paper uses the K-means clustering algorithm and the Gaussian mixture model (GMM) algorithm to establish two clustering algorithm models to identify the financial whitening degree.In order to verify The effectiveness of the algorithm identification model of the financial whitening degree clustering constructed by using the financial indicators after data normalization . The model creation process in this paper is as follows:

First using the StandardScaler function of sklearn to complete the normalization of 14008 sample data, and then use the K-means object of sklearn to create a K-means algorithm model framework. Set the parameter n_clusters of the number of cluster centroids that comes with the framework, that is, the K value, to 3; since the K-means algorithm is a locally optimal iterative algorithm that is sensitive to the initial value, set the number of randomly generated centroid arrays to 100. That is, 100 sets of initial centroid arrays are randomly generated, and each initial centroid array contains three random data coordinates as the initial centroids. Then run the model. Finally, the K-means algorithm model with the best comprehensive performance is obtained.

In addition, the GaussianMixture object of sklearn is also used to create a Gaussian mixture model framework, the cluster centroid number parameter n_clusters that comes with the framework is set to 3, the number of randomly generated centroid arrays is set to 100, and the number of model iterations Set to 1000. Then run the model to obtain the GMM algorithm model with the best overall performance.

The relevant evaluation indicators of the two models are shown in Table 3:

Table 3. Comparison of various performance evaluation indicators between the K-means model and the Gaussian mixture model

	Adjusted Rand index(ARI)	Mutual Information based scores(MI)	Fowlkes-Mallows scores(FMI)	Silhouette Coefficient	DBI
K-means	1.0	1.0	1.0	0.793	0.109

GMM	0.538	0.645	0.730	0.540	4.972

According to various test indicators, it can be seen that the performance effect of the relatively simple K-means algorithm model is very good, and its clustering results are completely consistent with the category results marked by the Excess-MAD algorithm in advance. Although the effect of the slightly more complex standard Gaussian mixture model is worse than that of the K-means model, it also has relatively good accuracy and has certain use value.

Then, based on the results of the GMM algorithm, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to judge the clustering effect under different initial cluster centroid numbers. The result is shown in Figure 3.

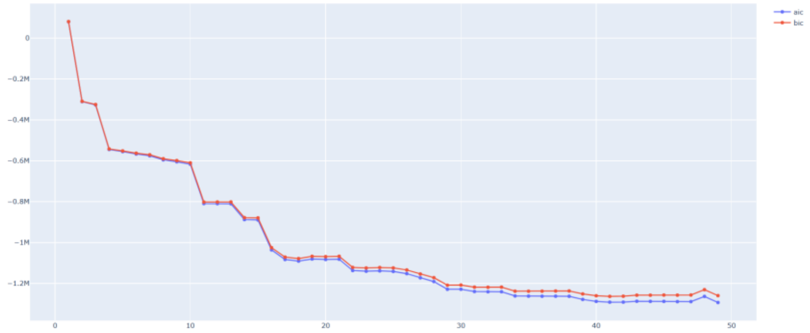


Fig. 3. Analysis results of Akaike information criterion and Bayesian information criterion based on GMM

The analysis results show that the total number of clustering categories starts from 2, and its AIC and BIC values are both negative numbers with large absolute values. However, the GMM with K=3 still has excellent clustering performance. And referring to relevant financial management theories, it can be seen that when the K value is 3, the model has the best interpretability. In addition, considering that GMM is more commonly used for generating density estimators and other applications than for generating clustering algorithm models. Therefore, it can be considered that when it is necessary to analyze a new set of company report data objects, the trained GMM can be used to roughly estimate the distribution probability density to assist users in making manual judgments in combination with relevant financial theories, thereby getting higher interpretability at the expense of accuracy ; or directly use the trained K-means model for recognition and judgment with poor interpretability but high accuracy. In summary, the two cluster recognition models obtained in this study based on financial indicators training can meet the needs of different usage scenarios, that is, they can meet their respective accuracy levels and understandability in their respective best usage scenarios and the goal of better judging the degree of financial whitewashing of corporate financial statement data is finally achieved. Therefore, the model trained in this paper has certain application value, which is also the greatest significance of this paper.

5 Research conclusion

This paper takes all listed companies in the chainese mainland capital market from 2000 to 2021 as the research object, and uses the Excess-MAD algorithm to initially identify the degree of whitewashing of corporate financial statements, and then selects financial statement data that account for about 15% of the total, after theoretical screening and analysis. After the Kendall correlation coefficient was screened, two clustering algorithm identification models were established after normalization. The empirical results show that the two clustering algorithm models obtained through training have sufficient recognition accuracy, and can respectively meet the needs of two different types of users in two different usage scenarios. In addition, compared with the direct use of the Excess-MAD algorithm based on the improved Benford's law, the application of the trained clustering algorithm model for report evaluation has the advantages of less input data, shorter judgment time, more in line with the classical financial theoretical analysis framework, A series of advantages such as high comprehensibility of model running result data. And it also shows that when studying the degree of whitewashing of corporate financial statements, it is not only feasible but also effective to use clustering algorithms represented by K-means algorithm and GMM algorithm to build models. Since the practical research in the field of intelligent financial analysis, especially in the identification of corporate financial statements is still relatively weak in china at this stage, this study also has a certain practical value.

In the next step of research, researchers can further explore on the basis of improving the operating efficiency of the model, and further improve the accuracy of the model to determine the data in the middle area by increasing the model entry indicators that fit the relevant financial theoretical framework, thereby improving the practicability of the GMM model. In addition, it is also possible to explore the specific application of improved algorithms from these two types of basic models to derive more and better new models, such as using the ARSGOS algorithm to improve the clustering performance of the GMM model on unbalanced data sets. The above-mentioned technical comprehensive improvements to the clustering algorithm model are also valuable research directions for further research.

References

[1] Chen Hanwen, Ding Peng.:Discussion on Several Issues of Financial Whitewashing . Chinese Agricultural Accounting.6-7(1996)

[2] Barney, Bradley J.: Schulzke, Kurt S. Moderating “Cry Wolf” Events with Excess MAD in Benford's Law Research and Practice. Journal of Forensic Accounting Research, 1(1). 66–90 (2016).

[3] Cong M , Li C , Ma B Q . :First digit law from Laplace transform. Physics Letters A.(2019)

[4] Hal V. Benford’slaw.American Statistician.65(1972)

[5] CarslawC. :Anomalies in Income Numbers:Evidence of Goal Oriented Behavior.The Accounting Review. 321-327(1988):

[6] Feng Yu, Ding Guoyong. :Banford's Law and Its Auditing Application. Auditing:theory & Practice. 44-45(2003):

[7] Chen Xi, Wan Yufei, Li Lu.: The Applicability of Finding Corporate Fraud Based on Benford's Law——An Empirical Test on the Financial Data of my country's Listed Companies. Finance and Accounting Monthly. 45-48(2012)

[8] Macqueen J. :Some Methods for Classification and Analysis of MultiVariate Observations. Proc. of, Berkeley Symposium on Mathematical Statistics and Probability. 281-297(1967)

[9] Wu Shisong, Wang Jianyong, Yan Yuping.: Simulation Study on Optimal Identification of False Information by Enterprises Computer Simulation. 313-316(2017)

[10] Guo Yang, Kong Wenjia, Feng Hening, et al.: Big Data Mining Applied to the Construction of Electric Power Enterprise Evaluation System . Mathematics in Practice and Theory.117-123(2019)

[11] Shen Ruyuan, Pang Deliang. :Corporate Credit Evaluation Based on Regional Spatial Relevance Characteristics: A Case Study of Listed Companies in Three Western Provinces. Journal of Yunnan University of Finance and Economics.139-145(2014)

[12] Bao Xinzhong.: Financial early warning based on particle swarm optimization K-means algorithm and rough set theory. Journal of Systems & Management.461-469(2012)

[13] Cao Peng, Li Wei, Zhao Dazhen. :ARSGOS Algorithm for Unbalanced Datasets . Journal of Chinese Computer Systems.818-823(2014)

[14] Wang Jingyong, Wang Yuanchang. :Can Earnings Management Increase the Predictive Ability of Financial Crisis Warning? Evidence from ST Listed Companies in China. Journal of Yunnan Normal University(Humanities and Social Sciences Edition).133-141(2010)

[15] Benford F.:The law of anomalous numbers. Proceedings of the American Philosophical Society.551-572(1938)

[16] Wang Jiamin, He Ding. :Financial Fraud Risk Governance Effect of Actively Shorting Chinese Concept Stocks——A Case Study Based on Benford's Law Compliance Test. The Chinese Certified Public Accountant.42-46(2022)