The Application of Cluster Analysis Technology in the Financial Field

Zhang Xian，Sun Baoshan

***Abstract*—Data mining refers to the process of searching the massive data stored in the data warehouse through related algorithms and obtaining the hidden information.Cluster analysis is one of the main technologies in data mining, which is widely used in the field of financial investment.This paper takes similarity as evaluation standard, uses cluster analysis method to detect and analyze various stock indicators, and provides corresponding investment suggestions according to the actual situation.Experiments show that this method can help users to make reasonable prediction of the growth trend of stocks, and finally assist them to make the right investment decision.The experimental results show that this method is more convenient and effective than other methods.**

***Index Terms*—data mining;Cluster analysis;Financial industry;Investment analysis**

I. Introduction

With the development of The Times and the progress of science and technology,people's ability to use computer technology to obtain data and information has been fully improved,but it also brings new challenges. Firstly,we find that data surplus has become a problem that has to be solved at present; Secondly,the lack of information is always troubling all kinds of enterprises.

With the help of data mining technology,more interesting information can be provided to the system database.People usually come into contact with a lot of information in the process of working in financial institutions. It is convenient to process the data through the database,but it is difficult to find the correlation between the data,so it is impossible to use it to make predictions about the future situation.If this information is put to use,it will help us find out by learning the internal principles and algorithms of data mining.

The interdependence and law between data,and further reduce the financial risk of enterprises.In data mining technology,clustering technology can help users quickly control and grasp various characteristics of stocks,analyze the growth ability of stocks,and help investors develop more perfect investment strategies.

II. Cluster analysis technology

A. Basic Concepts

Clustering is the process of dividing data into different clusters based on its different characteristics.The goal is to make the distances between individuals belonging to the same category as small as possible,and the distances between individuals on different categories as large as possible.The clustering method includes statistical method,machine learning method,neural network method and database oriented method.

According to the statistics, the multivariate data analysis is divided into three categories, and the cluster is among them. Euclidean distance, Minkowski distance and so on are the contents of clustering. Joining method, systematic clustering method, dynamic clustering method, decomposition method, fuzzy clustering, orderly sample clustering and overlapping clustering all belong to the traditional statistical clustering analysis methods. It uses the overall comparison method to check all the objectives to achieve the class division, so all the information must be given in advance, and adding new data objects in real time is not allowed. When the database is relatively large, the analysis difficulty also rises sharply, which is not suitable to use this method again.

B. Cluster Analysis Method

Cluster analysis is to split and classify the data set, so that the data can have more similarity in the same set, while in different sets, the data are very different. Using the clustering method, the similarity between the data can be calculated, and the distribution mode and the connection between the properties can be found out. The basic idea of cluster analysis is to find out the elements that can measure the similarity of the data in the investigated data set and compare the characteristics, and classify some characteristics into the same category; other differences are classified into other types. That is, the similarity of data objects is high in the same set, but not the same in different sets. Similar or different descriptions are determined based on the numerical values of the data description properties. Generally, the distances between different data objects are used for expression. Distance is the measure of similarity between samples and the basis for cluster analysis. Its distance must meet the following conditions:

(1) Non-negativity: for the arbitrary

$x_i, x_{\mathrm{j}}, d\left(x_{\mathrm{i}}, x_j\right) \geq 0$,If and only if i=j,. $d\left(x_i, x_j\right) = 0$.

(2)Symmetry: for the arbitrary.

$x_i, x_{\mathrm{j}}, d\left(x_i, x_j\right) = d\left(x_i, x_{\mathrm{j}}\right)$.

(3) Triangle inequality: a sum for all of them

$$x_i, x_j, x_k, d\left(x_i, x_j\right) \leq d\left(x_j, x_k\right) + d\left(x_j, x_k\right)$$

$$x_i = \left(x_{\mathrm{i}1}, x_{\mathrm{i}2} \ldots\ldots\ldots x_{\mathrm{i}p}\right)$$

$$x_j = \left(x_{j1}, x_{j2}, \ldots\ldots\ldots\ldots x_{jp}\right)$$

Is two P-dimensional data objects with common distances as follows:

Absolute value distance (Manhattan distance)

$$d\left(x_i,x_j\right)==\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$$

Euclid Distance (European type)

$$d\left(x_i,x_j\right)=\sqrt{\sum_{k=1}^{p}\left(x_{jk},x_{jk}\right)^2}$$

Minkowski Distance (Minkowski)

$$d\left(x_i,x_j\right)=\left(\sum_{k=1}^{p}\left(x_{ik},x_{jk}\right)^q\right)^{1/q}$$

Chebyshev Distance (Chebyshev)

$$d\left(x_i,x_j\right)=\max_k\left|x_{ik}-x_{jk}\right|$$

Mahalanobis Distance (Ma)

$$d^2\left(x_i,x_j\right)=\left(x_i-x_j\right)^T\wedge^{-1}\left(x_i-x_j\right)$$

Here is the covariance matrix of the random variables. At this point, the absolute value distance, Euclid distance, Chebyshev distance are respectively Minkowski distance in q=1,2,3... A special case of when, for general Euro, Chullidean spaces, and Mahalanobis distance for the special case as sampling points of random variables. Distance stances applies to distance measures as sample points of random variables. Can be achieved by using the SPSS. SAS software to complete the computational process of the cluster analysis.

C. Selection Of Cluster Analysis Index System

By investigating the basic status of financial enterprises, we find that the amount of profit and the speed of growth rate are the important concerns to measure the stock price, and also become the fundamental factor to determine whether the enterprise is the next investment target, the other time. The return on equity, main income growth rate and earnings per share are closely related to stock value to analyze the role of enterprise behavior on stocks.

1) Profitability indicator:

$$\text{margin on total assets}=\frac{\text{net margin}}{\text{average total assets}}$$

The overall profitability of the company depends on that.

$$\text{net worth}=\frac{\text{net margin}}{\text{average net assets}}$$

It reflects the size of the shareholders' return for their investment.

$$\text{major business profits}=\frac{\text{main business yield rate}}{\text{main business income}}$$

The main business is the main target of the main profit of the enterprise. The higher the yield of the main business, the easier the enterprise to take advantage of the market competition.

$$\text{earnings per share}=\frac{\text{net margin}}{\text{totalcapital stock at the end of the term}}$$

The higher earnings per share, the higher investment profits in the sector and the higher earnings per stock.

2) Growth ability index:

The growth performance of the listed companies with outstanding business performance will also be more obvious, and the profitability of the listed companies with excellent growth performance will also be stronger.

$$\text{main business revenue growth}=\frac{\text{current main business income}}{\text{main business income of the previous per iod}}-1$$

It represents the size of the enterprise market, the ability to expand, and shows the growth of the enterprise's main business.

$$\textit{net profit growth rate}=\frac{\text{current net profit}}{\text{net profit of previous period}}-1$$

The increase in net profit represents the company's capabilities, growth prospects and value to investors.

The profitability and growth indicators mentioned above are both positive indicators.

3) Capital capital expansion capacity index:

As a negative index, the listed companies with small share capital often have strong expansion ability and large growth space.

$$\text{net asset value per share}=\frac{\textit{ending net assets}}{\textit{ending capital stock}}$$

It is the amount of shareholders' equity represented by each share and represents the minimum intrinsic value of each common share.

III. APPLICATION INSTANCES

We studied 30 shares listed in Shenzhen and randomly selected 30 stocks. The index data of the sample stocks comes from the recent financial data released by Guotai Junan Great Wisdom Securities Network.

1) Standardization of raw data

In this paper, the different dimensions and magnitude of the original data can be avoided to conduct cluster analysis and discriminant analysis.

$\overline{x}$, s_j and R_j is the sample mean and sample range and sample standard deviation of the j th indicator.

$$x=\frac{1}{n}\sum_{y=1}^{n}x_y \quad s_j=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_y-\vec{x}_i)^2} \quad R_y=\max\{x_y\}-\min\{x_n\}$$

Standard deviation standardization

$$x_y=\frac{x_i-\overline{x}_j}{s_y}\quad (\mathrm{i}=1,2\cdot n,\ j=1,2,\cdot p)$$

Very poor standardization

$$\mathbf{x}_{ij}=\frac{\mathbf{x}_{ij}-\overline{\mathbf{x}}_j}{R_j}\ (i=1,2,\cdot n;j=i,2,\cdot p)$$

2) Positive treatment of the reverse index

Outstanding share capital is inverse index (the unit takes 100 million shares), the absolute value of this index takes reciprocal, namely

$$x_{ij}=\frac{1}{\left|x_{ij}\right|}\quad (i=1,2,\cdot n:j=1,2,\cdot p)$$

IV. The experiment

Systematic clustering analysis of 30 stocks was performed using spss software.

The first category:Consensus Pharmaceutical,Great Wall Development, Shenchiwan B, Baoxin Software, China Software,Antai Group,New Huangpu,Xiangtan Electrochemical,Zhongbing Optoelectronics,Jincheng Shares, Fengfeng High-tech, Yaxing Bus.

The second category: Jiangsu Wuzhong, Fangda B, Public Technology, Xindu Hotel, North Mine Magnetic Materials, Tengda Construction, Huamao Shares, Jingxin Pharmaceutical.

The third category: Shanghai Pudong Development Bank, Sinolink Securities, Dun'an Environment, Public Technology, Guotong Pipe Industry, TCL Group, Handan Iron and Steel.

From the above classification results, we can know that:

The first type of stocks are obviously low returns, growth, low, is a poor performance of stocks. The relatively low earnings per share, net assets per share, especially the net assets per share significantly lower, equity expansion ability, it also represents the poor stability is the first kind of the shortcomings of listed companies, stagnant development, and the main business negative growth year-on-year growth rate and low net profit year-on-year growth rate can be seen that the company has insufficient growth. Therefore, it is judged that the enterprise is in the recession stage, and the investment value is relatively low.

The second type of stocks have good returns, high growth, but have low operating cash flow. It shows that the sales of listed companies are not smooth, and the capital turnover is slow, so such stocks should not be long-term investment.

The third type of stocks has rapid growth and smooth business, so this kind of enterprise has strong future development momentum and little risk. Investors should prioritize them as long-term targets.

V. Cluster effect test

Is the grouping obtained through the above clustering process satisfactory? After grouping, the differences between samples within groups should be small, while the differences between groups should be large. To verify the effect of clustering grouping above, it can be tested using ANOVA. An Analysis of Variance (ANOVA) is a test of the equality of multiple normal population means with one, two, or more factors. Analysis of variance (ANOVA) has been widely used in various analytical studies in different fields. Methods starting from the variance can help us to discover the inner laws of things.

We program it through the Matlab language, and the program uses the manova1 command. D = manova1 (X, group, alpha); in the input value at the right end, x is the observation data array, 13 rows represent n samples, and the Pcolumn represents Pvariables. A group is a vector of n row 1 columns, which records the corresponding scores of n samples X and gr o up are data that must be entered using this procedure. The Alpha is the significant level. The study null hypothesis needs to be tested in this 10-dimensional space, so that there is no significant difference between each population mean value. Calculated d1, representing the null hypothesis, was rejected and constituted significant differences between groups in 2-dimensional space. Representative cluster analysis results are credible at this time.

VI. Conclusion

Cluster analysis Through clustering, the growth of listed companies is analyzed through comprehensive indicators, and quantitative analysis of the clustering results is conducted to obtain the strength of each company. Based on the fundamental quantitative analysis, the study of the intrinsic value of stocks is conducive to investors to narrow the scope of investment choice, determine the investment value, reduce the investment risk, and provide reference for managers.

The feasibility of the experimental results was further verified by applying the cluster analysis technique and using multivariate analysis of variance.

Acknowledgment

Research on the Chinese Academic Degrees and Graduate Education Association

References

[1] Fan Zuojun, Guan Wei. A partitioning method of differentiated regional financial regulation- -the application of system-based clustering analysis method [J]. Managing World, 2008 (4): 36-47

[2] Shen Tao, Liang Haisen. Evaluation of financial ecological environme nt in ASEAN countries based on cluster analysis [J]. Social Science, Guan gxi, 2016,0 (10): 41-45

[3] Tian Lin. Clustering analysis of regional financial comprehensive competitiveness and optimization and integration of financial resources [J]. Financial Theory and Practice, 2006 (2): 16-18

[4] Zhang Xian. Data mining technology and its application in the financial field [J]. Finance Teaching and Research, 2003 (4): 15-18

[5] Chai Xiaohui. Application and development trend of data mining in thefinancial field [J]. South China Financial Computer, 2004,12 (9): 97-99

ZhangXian Male,Shijiazhuang,Hebei Province,Graduate Student,Tianjin University of Technology school of Computer Science and Technology, Tianjin Xiqing District Guest Water, 399 West Road, 300387, China.

Sun Baoshan Male, Doctor, Master's Supervisor, Member of CCF, Wit h research interests in natural language processing, Big data analysis, Int elligent algorithms, and embedded system design Development