Application of Fuzzy C-Means Clustering and Support Vector Machine in Stock Price Analysis

Jinliang Wang[1,2], Wennan Wang[1], Tuli Chen[1], Fu Luo[1], and Shiyang Song[3(✉)]

[1] School of Management, Guangdong University of Science and Technology, Dongguan, China
B21092100207@cityu.mo
[2] DBA Candidate, Faculty of Business, City University of Macau, Macau, China
[3] Alibaba Cloud Big Data Application College, Zhuhai College of Science and Technology, Zhuhai, China
2463756962@qq.com

Abstract. With the rapid development of the global economy and the continuous expansion of the investment scale in the financial market, more and more transaction data and market public opinion information are generated in the stock market under the background of big data, which makes it more difficult for investors to distinguish effective investment information. This paper presents a stock price prediction method based on fuzzy clustering and support vector machine. Fuzzy clustering has the characteristics of high accuracy when processing large data. When analyzing the financial information of listed companies, fuzzy clustering technology and related index method can effectively reduce the error. Through the analysis of the factors influencing stock value investment, this paper selects five aspects from the financial statements of listed companies that can reflect their profitability, development ability, shareholders' profitability, solvency and management ability. This paper pays attention to the verification of the theoretical method model, using fuzzy clustering, support vector machine and bp neural network to compare the data, to ensure the effectiveness of its practical application. In this paper, the real data of China's stock market are used for testing. The accuracy and recall rate of mohujulei model are relatively stable, with the accuracy of 0.884 and 0.001 respectively. The research of this paper is helpful to improve the quantity and quality of listed companies.

Keywords: Fuzzy Clustering Algorithm · Correlation Index Method · Support vector machine · Stock Price · Price Prediction

1 Introduction

There is room for arbitrage in investing in the stock market. High profits also bring high risks. As a result, investors always try to determine and estimate stock values before taking any action, but stock values are often affected by economic and external factors beyond their control, which makes it very difficult to identify future stock market trends. Therefore, the traditional forecasting model is not enough to predict stock

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 800–807, 2023.
https://doi.org/10.2991/978-94-6463-198-2_83

volatility. Looking back at the development of financial markets, from the New York Stock Exchange in 1929, the positive and negative "chain reaction" caused by ultra-high leverage made the stock market continue to plunge after a wild rally. On October 19, 1987, the "Black Monday" stock market crash in the United States sent the Dow Jones Industrial Average plunging 20 percent in a matter of hours, triggering panic in financial markets. So researchers are constantly experimenting with new models of stock returns to see if they can predict future market returns. Fuzzy clustering and genetic algorithm are widely used data mining tools in the financial field in recent years. It has the characteristics of high efficiency and small information loss to deal with the large scale database with various data attributes [1]. Fuzzy clustering technology and correlation index method can effectively reduce the massive financial fundamentals information of listed companies. Support vector machines (SVM) have gained popularity in machine learning algorithms, often used to predict stock prices and optimize stock market predictions using the core parameters of SVM [2]. This paper will use the software for principal component analysis, the selected stock related indicators into internal factors or external factors, and then use the fuzzy clustering method and deep neural network to effectively analyze the financial information of listed companies. For the analysis of extreme risks, this paper will use the deep neural network and fuzzy clustering extreme risk warning model of stock market to apply to the real 300 index of China to conduct the early warning analysis of extreme risks.

2 Research Methods

2.1 Extracting Feature Indexes

Principal Component Analysis (PCA) is used to extract the characteristic indicators. Principal component analysis (PCA) has been widely used in the fields of demography, biology, psychology or genetics. But research into its use in finance is relatively recent and remains rare, especially in the context of portfolio management. Principal component analysis (PCA) is actually a statistical method, which can reduce the size of data set without losing important information so as to achieve dimensional reduction of data set. From the perspective of mathematics, Sharma (1996) proposed that principal component analysis is a technology that uses new variables as linear combinations of original variables to form new variables, To maximize variance, the two variables should be orthogonal to each other. For the purpose of obtaining the comprehensive score of the evaluation index, we used principal component analysis to analyze the main financial index data of some listed companies. A simultaneous analysis of financial index data and transaction index data was conducted in order to predict stock prices [3].

Principal component analysis was used to separate the indexes into two groups with high collinearity based on internal and external factors. As an extension of the integrated moving average model, the RAROC model gradually becomes stable as the phase difference of the non-stationary sequence increases, and the newly obtained stationary sequence can be modeled with the RAROC model and the original sequence can be obtained by applying the inverse transformation.

2.2 Selection of Measurement Indicators

Various index data centers published by listed companies can be used to analyze the company's operation, and some data can be used to measure the return on investment, etc. The financial risk assessment index system with four dimensions of solvency, operating ability, profitability, growth ability and cash flow ability can affect the financial risk of enterprises [4].

2.3 Stock Value Feature Selection Using the Fuzzy Clustering Algorithm

Fuzzy clustering This method can find overlapping clusters, but need to maintain and calculate the data size of a member matrix multiplied by the number of clusters. The traditional classification method uses statistical methods to cluster similar and homogeneous data, but the data in the financial stock market is diversified and there are many classification standards. The direct selection of indicators by qualitative analysis method is likely to lead to a large amount of information missing, which will directly affect the accuracy of investment decisions[6]. The selection of criteria should enable the data to be categorized into different categories. However, in practical applications, the samples may belong to different clusters and have different degrees of membership. Fuzzy clustering uses fuzzy logic to allocate data to different clusters, which provides an effective solution for separating overlapping clusters. By calculating the reverse distance to the center of the cluster, the correlation with the cluster is verified. The clustering center of fuzzy clustering verification obviously depends on the geometric position of data points in the plane or space. In the fuzzy clustering algorithm, an objective function that needs to be

$$F(U, V, m; X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (u_{ij}) \left\| x_j - \upsilon_i^2 \right\| \tag{1}$$

where m is the fuzzy factor, k is the number of clusters, $V = (v_1, v_2, ..., v_k)^T$ is the vector of cluster centers containing k cluster centers, n is the number of data points, and $X = (x_1, x_2, ..., x_n)^T$ is the vector of data points. $U = \left[u_{ij}\right]_{k \times n}$ is the matrix of affiliation containing the affiliation u_{ij}. Denotes the degree of affiliation of x_j in the i-th cluster. $\|.\|$ denotes the Euclidean distance standard ($\|Z\| = \sqrt{Z^T \cdot Z}$). m is used to normalize and fuzzify the membership relationship, and their sum should be equal to 1. Minimizing $F(U, V, m; X)$ is achieved by iterative techniques such as alternating optimization.

2.4 Standardization of Data

As a result of existing research methods, data standardization is applied in this chapter, and the specific algorithm involves two steps [5].

① Translation-standard deviation transformation:

$$x = \frac{s - x_k}{s} (k = 1,2, \ldots ,n) \tag{2}$$

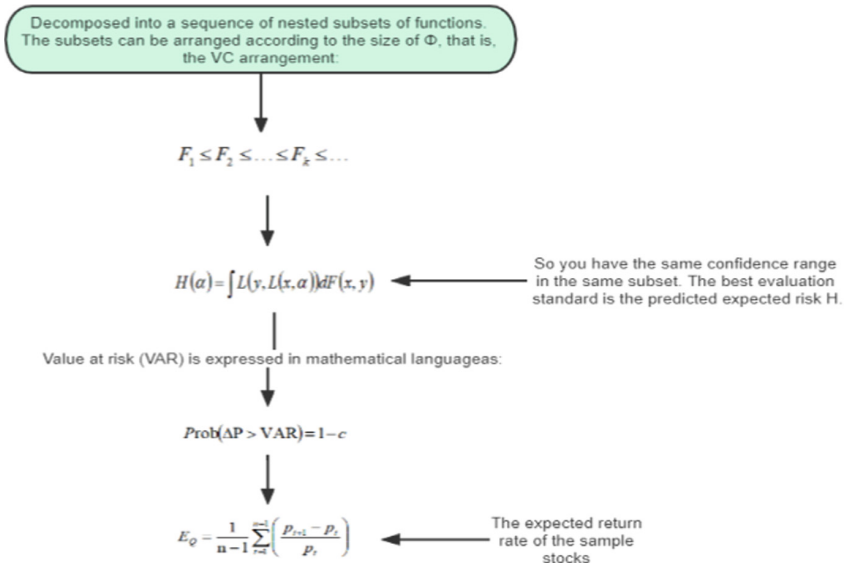

Fig. 1. Model building Process

Among them,

$$x_k = \frac{1}{m}\sum_{i=1}^{m} x \tag{3}$$

After transformation, each variable has a mean value of 0 and a standard deviation of 1, so the data have no dimension. The interval x_k is not guaranteed, however.

② Translation-Polarity Transformation.

$$x'' = \frac{x' - min\{x_k\}}{max\{x_k\} - min\{x_k\}} \tag{4}$$

It is clear that all x'' are on the interval [0, 1], and the effect of the gauge is eliminated.

A correlation coefficient is first calculated between each indicator, followed by calculating the average square of the correlation coefficient between each indicator and other indicators. To measure the correlation, choose the largest index. The indicator set can be generated automatically if only one indicator exists in the classification. Select either indicator if there are two in the classification.

2.5 Stock Price Prediction

When predicting stock prices using empirical risk minimization principles, the traditional machine learning method is limited by the number of learning samples and the effect of confidence intervals. Stock prediction behavior is often characterized by poor data and training samples due to the few known samples obtained. Also, low correlations among

the samples. Therefore, we need to find another method to try to optimize this problem. Figure 1 shows the calculation process of stock price prediction proposed in this paper:

As a result of determining the correspondence between the inputs and outputs of a known training sample, a joint probability, F(x,y), can be expressed for the unpredicted variables x and y. If there are no correlations between the inputs and outputs, then all n samples are identically distributed and independent of each other. In order to determine the expected risk, one must estimate the function F(x,y) using the optimal solution to the prediction function set $\{f(x, w)\}$.

3 Results

Compared with R1 and R2, the stock selection model proposed in this paper achieves better results in terms of average return and cumulative net worth. On the one hand, the model exceeded R1's average return rate in seven of the nine test cycles and R2's average return rate in eight of the nine test cycles. In addition, the model achieved positive returns in all test periods except the third quarter of 2019 and the first quarter of 2020, which may have been affected by the stock market crash. In addition, the cumulative net worth of this model was higher than R1 and R2 in all test periods. The results show that the stock selection model proposed in this paper is an effective investment strategy. A number of other advantages were achieved by the SVR Max and Min stock selection models in terms of AR, Sharpe ratio, Probability.(R1), Probability.(R2), and Hit. Stock selection models perform better than a back-propagation neural network in each evaluation metric. By doing so, the FCM provides market predictors that can be used to assist in selecting stocks. Based on the comparison of each benchmark model, the FCM outperformed the others significantly. Normality tests are required before performing a t-test.As shown in Fig. 2.

Results show that the fuzzy clustering method predicts stock prices more accurately and in less time than all benchmark models. There was a significant improvement in FCM ability to predict in this study. DM test is conducted for each benchmark model in this paper to demonstrate statistically that FCM is significantly better than other benchmark models. FCM has strong predictive capability. Stock selection decisions can be improved by using a predictor based on FCM structure. Five-fold cross-validation of the SVM model generally yields reliable results. Model performance is excellent, as

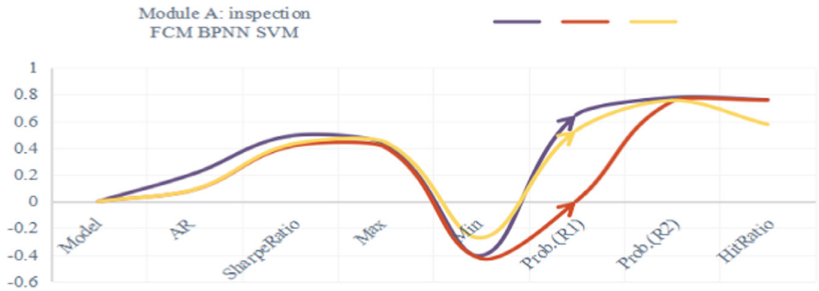

Fig. 2. Decision contrast in Module A

accuracy and recall values are very consistent. Figure 3 shows the SVM model's overall inspection and comparison.

As a result of significant market volatility in the third quarter of 2019, there was an increased risk of market risk. SVM and FCM issued crisis warnings during this time period. The model warned about extreme risks before the "stock market crash." Nonetheless, SVM and FCM are still susceptible to market risks. In crisis states, support vector machines were accurate to 0.877, and FCM models were accurate to 0.831. SVM and FCM both had an overall accuracy of 0.722 and 0.821, respectively. In accordance with the theoretical analysis, the results support the conclusions. The support vector machine (SVM) model predicts crisis samples positively, but produces more false alarms than the other models. In general, FCM is a sound model, but it does not respond well to crises. Both SVMs and FCMs have produced favorable predictions for extreme risk early warning, there are certain similarities and complementary features between each model. As shown in Fig. 4.

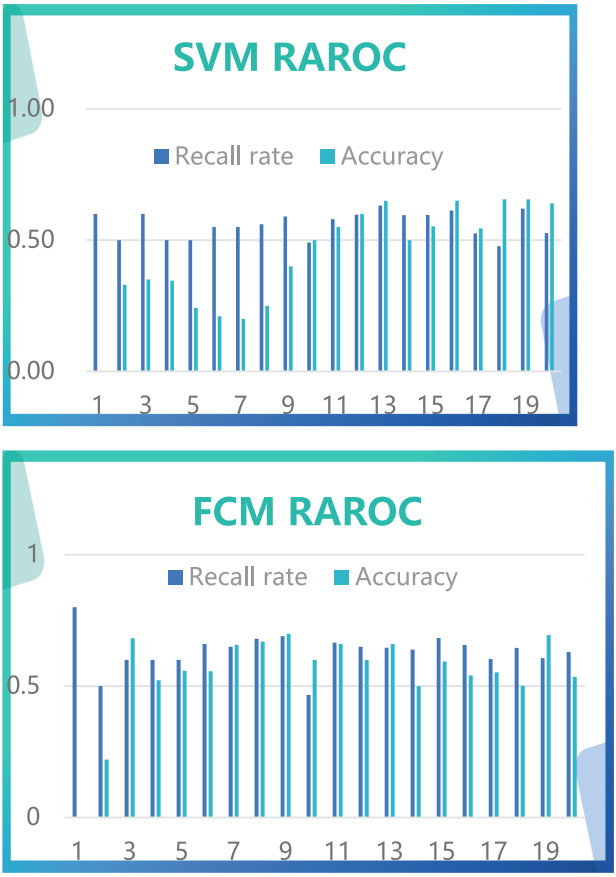

Fig. 3. SVM model overall test comparison

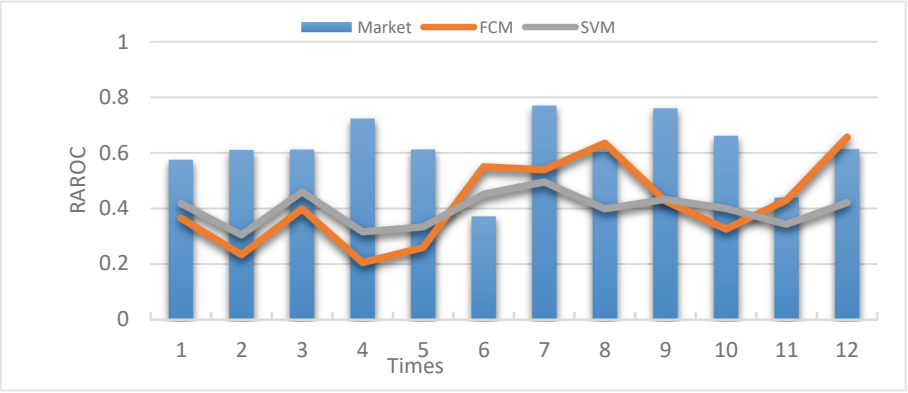

Fig. 4. Performance comparison of the extreme risk early warning model

4 Discussion

According to the experimental results, it can be found that the fuzzy clustering and deep neural network model can play an early warning role in the extreme risk of the stock market to a certain extent. In the case of large fluctuations in the stock market.

As can be seen from the prediction accuracy results of the model, both the deep neural network and the fuzzy clustering model have achieved good prediction effect in the extreme risk warning model of the financial market. Although there are some errors in the prediction effect of the deep neural network algorithm on the extreme risk, resulting in crisis misjudgment, the overall prediction effect is relatively good. The fuzzy clustering model is better for extreme risk warning model. The fuzzy clustering model has good prediction effect, but it is not sensitive to crisis signal. Although the two models have their own shortcomings, they can complement each other to achieve more accurate prediction results in actual risk prediction.

5 Conclusion

The optimization effect of fuzzy clustering in the multi-factor stock selection model is much better than that of the support vector machine algorithm. When using the fuzzy clustering method to screen the influence factors, the use of the influence factors of stock selection investment can also obviously achieve better cumulative excess return. Through comparison, it is found that the excess return rate of the stock selection model considering only financial factors is much lower than that of the stock selection model considering both financial factors and prediction factors. The multi-factor stock selection model and prediction factor have complementary functions, and the prediction factor has a positive effect on the multi-factor stock selection model. At the same time, from the inspection data of the portfolio data, the weight coefficient of the prediction factor proposed in this paper is significantly higher than that of other financial factors.

Acknowledgments. Guangdong Province key construction discipline scientific research capacity improvement project: Guangdong Province manufacturing industry digital transformation research

(2021ZDJS113S); University level collaborative innovation center of Guangdong University of science and technology: Collaborative Innovation Center for e-commerce and logistics application talents training (GKY-2019CQYJ-8).

References

1. CAO, Jiasheng; WANG, Jinghan. Exploration of stock index change prediction model based on the combination of principal component analysis and artificial neural network. Soft Computing, 2020, 24.11: 7851-7860.
2. ENKE, David; MEHDIYEV, Nijat. Stock market prediction using a combination of stepwise regression analysis, differential evolution-based fuzzy clustering, and a fuzzy inference neural network. Intelligent Automation & Soft Computing, 2013, 19.4: 636-648.
3. GAO, Bo. The use of machine learning combined with data mining technology in financial risk prevention. Computational Economics, 2022, 59.4: 1385-1405.
4. Li Xiuge. Research on fuzzy clustering theory based on fuzzy equivalent matrix. 2015. PhD Thesis. Shenyang: Liaoning University.
5. NTI, Isaac Kofi; ADEKOYA, Adebayo Felix; WEYORI, Benjamin Asubam. Efficient stock-market prediction using ensemble support vector machine. Open Computer Science, 2020, 10.1: 153-163.
6. WANG, Wennan, et al. Stock price prediction methods based on FCM and SVM Algorithms. Mobile Information Systems, 2021, 2021.