Research on Asset Allocation Based on Hierarchical Clustering and Fuzzy Theory

Lirui Xiao[1], Xue Deng[2,*]

[1]School of Economics and Finance, South China University of Technology, Guangzhou, Guangdong, China
[2]School of Mathematics, South China University of Technology, Guangzhou, Guangdong, China
**Corresponding author*

Abstract: *After Markowitz's portfolio theory was put forward in the 1950s, the optimization method of venture capital portfolio had made continuous progress in three branches. Among them, double objective function became a mainstream. On the basis of predecessors, this paper proposes an asset-weight-optimization method based on hierarchical clustering and fuzzy theory. Firstly, this paper extracts the four characteristics of stocks, uses the hierarchical clustering method to aggregate the stocks with strong correlation to form a group, and then selects two of the most representative stocks from the group (i.e. clustering the data points with the best performance in each group) to form a new stock portfolio. Then, the fuzzy programming of the expected return is carried out for the new portfolio. And finally, the optimized asset allocation weight is obtained. In the experiment, we use the data from CSMAR database for empirical analysis. For the data preprocessing process of hierarchical clustering, we choose Ward algorithm. We conduct experiments with stock pools in China with data sample sizes of 20, 35 and 45 respectively. Firstly, we use hierarchical clustering data preprocessing, then optimize to obtain the weight with fuzzy theory, and finally compare the performance of relevant indicators with different sample sizes. Experiments show that hierarchical clustering can reduce the dimension of stock correlation matrix, which breaks through the limitation of the number of stocks for experiments, realizing "large-scale stock selection". The performance of programming which uses hierarchical clustering and fuzzy theory method is better than direct linear programming, in terms of simulating the state of people's investment and meet people's expectations at the same time. The innovation of the method and the improvement of the performance of relevant index show that the method proposed in this paper has a certain reference.*

Keywords: *Hierarchical; clustering and fuzzy theory, asset allocation, Programming, stocks*

1. Introduction

Markowitz published *Portfolio Selection* [1] in 1952, which opened the prelude to "Quantitative Investment". In that paper, Markowitz mainly studies how people make decentralized investment to maximize investment returns when facing systematic and non-systematic risks in investment. For more than half a century, people have been making in-depth exploration based on Markowitz research, which makes the development of this theory more and more perfect.

In exploring the allocation of risk portfolio weight, there are three mainstream directions. The first direction is to take the expected return as the objective function for optimization. In this direction, the methods of "Expected Utility", "Prospect Theory" and "Minimax Regret Theory" were introduced. They describe the reality closer and closer. The second direction is to minimize the risk. At first, it was the decentralization theory, and then ideas of "Minimum Variance", "Minimum Semi Absolute Deviation" [2] and "Entropy Constraint" [3] were introduced. The third direction is to consider the optimization of income and risk at the same time, and use the double objective function to solve the optimization. However, the methods above have some shortcomings. The selection of weight depends on the value of correlation matrix, but the correlation matrix still lacks the concept of hierarchy, which allows the weight to change freely in an undesirable way. For example, because the number of samples increases with the square of the matrix, it also has the disadvantage of very small correlation; Without considering that stocks of similar nature can be merged to reduce the risk; It is usually the best within the sample, but it performs poorly outside the sample. Raffinot [4] proposed the idea of introducing “Hierarchical Clustering” to financial markets in 2017. The inspiration comes from complex systems such as financial markets that often have a certain structure and are usually organized in a hierarchical way. Therefore, hierarchical clustering not only solves the problem that the correlation matrix of stock portfolio is too complex, but also reduces the risk of portfolio.

In addition, in the actual investment environment, many constraints cannot get accurate figures, but can only get a certain range of degree. At the beginning of the 21st century, Tanaka and Guo [5,6] proposed a portfolio model with fuzzy number return distribution. Its central idea is based on probability theory and mean variance model. In 2001, Carlsson and Fuller [7] defined the possible mean and possible variance of fuzzy numbers. Then, Carlsson [8] and others introduced a portfolio possibility method with the most efficient use score. These methods can fit the real situation better than a single number, and the results are more accurate.

This paper selects a vertical analysis method, combing these two fields together. Firstly, hierarchical clustering is used to reduce the dimension of matrixes and reduce the risk at the same time. Then, the double objective function is characterized by fuzzy theory, and then it is optimized by programming.

In evaluating the effectiveness of portfolios, there are often the following ways: Markowitz's "Mean Variance" portfolio theory [1]; Sharp's "Sharp ratio" theory. In addition, Fang [9] used the cross DEA evaluation to evaluate the combination efficiency, and Li et al [10] proposed the evaluation method based on TOPSIS multi-attribute decision-making. This paper uses the "decision satisfaction" score obtained by the double objective function as a reference for evaluating the portfolio.

2. Concept, formula and model

2.1. Clustering feature selection for stocks

This paper selects four common characteristics of stocks for clustering: the average monthly return (considering cash dividends), variance calculated from monthly returns, Amihud indicators, Book-to-market ratio during the reporting period.

Liquidity refers to the ease with which an asset can be successfully realized in the market at a reasonable price. It is very important for a stock. First, liquidity reflects the risk of a stock. Second, it reflects the public's attention to a stock, which has a strong correlation with the performance of the stock price.

This paper uses Amihud index to measure liquidity. Amihud index is the ratio of stock return and trading volume over a period of time, which measures the sensitivity of stock price to trading volume: if the change of stock trading volume will lead to sharp fluctuation of stock price (sharp rise and fall), the larger the Amihud index, the worse the stock liquidity; On the contrary, if the change of trading volume has less impact on the change of stock price, it indicates that the liquidity of stock is better [11]. The calculation formula is as follows:

$$Amihud_{i,t} = \frac{1}{D_i} \sum_{d=1}^{Di} - \log\left(\frac{\left|R_{i,d}\right|}{\left|Volume_{i,d}\right|} \right) \tag{1}$$

Where, i represents the stock, t represents the year and d represents the trading day; R represents the daily return rate of stock (%); Volume represents the daily trading volume (one million yuan).

Book to market ratio is a mathematical comparison of a company's actual value and market value. The actual value of the company is determined by internal accounting, and its market value is its market value. Market analysts can use the results of this comparison to judge whether a company is overvalued or undervalued. The calculation formula of book to market ratio (BM) is:

$$BM = \text{Shareholders' equity} / \text{Company market value} \tag{2}$$

2.2. Hierarchical clustering

A dataset has the characteristics of multiple dimensions. Hierarchical clustering attempts to divide the data set at different levels according to the value of features, so as to form a tree like clustering structure. The data set division can adopt the "bottom-up" aggregation strategy or the "top-down" splitting strategy. Common hierarchical clustering includes Single-Linkage algorithm, Average-Linkage algorithm, Ward method (WM), Directed bubbling hierarchical tree method (DHBT), Birch algorithm, Rock algorithm, etc. This paper adopts Ward algorithm. Ward algorithm has two main characteristics: one is to deal with large data sets, and the other is the robustness to outliers.

2.2.1. Ward Algorithm

Ward algorithm is proposed by Joe H. Ward [12]. The distance between classes is expressed by cosine (cosine distance). Cosine distance is often used to express the similarity of two eigenvectors, and the value is [-1,1]. Assuming two N-dimensional eigenvectors X, Y, the remaining chord distances can be expressed as:

$$\frac{\sum_{i=1}^{N} X_i \times Y_i}{\sqrt{\sum_{i=2}^{N} \left(X_i\right)^2} \times \sqrt{\sum_{i=1}^{N} \left(Y_i\right)^2}} \tag{3}$$

Ward algorithm is based on a new formula error sum of squares (ESS) to calculate distance, and M_i are data points.

$$ESS = \sum_{i=1}^{n} \vec{M}_i^2 - \frac{1}{n} (\sum_{i=1}^{n} M_i)^2 \tag{4}$$

(1) The steps of Ward algorithm are as follows:

(2) Initializes each point into a cluster. Set ESS to 0 in each cluster.

(3) Calculates the total ESS of all clusters.

(4) If each cluster is to merge into a new cluster, calculate the ESS between them.

(5) Merges two clusters in a manner that minimizes the increase in the total ESS value.

(6) Repeat the above step until all clusters are merged into a cluster.

2.3. Intuitive triangle fuzzy number

If the fuzzy number $A = \left[l, m, u\right]$, where, $0 \le l \le m \le n$, then A is called a triangular fuzzy number. Its membership function $\mu_A\left(x\right)$ satisfies:

$$\mu_A\left(x\right) = \begin{cases} \dfrac{x-l}{m-l} \cdots x \in \left[l, m\right] \\ \dfrac{x-u}{m-u} \cdots x \in \left[m, u\right] \\ \\ 0 \cdots x \in othersets \end{cases} \tag{5}$$

Where l, u is the upper and lower bounds of A, and m is the median of A. In this paper, it is obtained that the return function of investors is the function of investment proportion. See formula (6) for details:

$$N\left(x\right) = \sum_{i=1}^{n} R_i x_i - \left[\sum_{i=1}^{n} k_i \left|x_i - x_i^0\right|\right]. \tag{6}$$

The membership functions of the two objectives of maximum return and minimum risk are expressed as follows:

$$\mu_{\max}\left(x\right) = \begin{cases} 1, N\left(x\right) \ge N_1 \\ \dfrac{N\left(x\right) - N_0}{N_1 - N_0}, N_0 < N\left(x\right) < N_1 \\ 0, \quad N\left(x\right) \le N_0 \end{cases} \tag{7}$$

$$\mu_{\min}\left(x\right)=\begin{cases} 1,\sum_{i=1}^{n+1} d_i x_i \le M_1 \\ \dfrac{M_0-\sum_{i=1}^{n} d_i x_i}{N_1-N_0}, M_1<\sum_{i=1}^{n+1} d_i x_i < M_0 \\ 0, \quad \sum_{i=1}^{n+1} d_i x_i \ge M_0 \end{cases} \tag{8}$$

Where, N_0, N_1 is the acceptance degree of the expected return given by the investor according to the subjective risk M_0, M_1 is the acceptance degree of expected risk given by investors subjectively, which can better reflect the needs of investors under different conditions. d_i is the mean deviation of risk asset i return.

Suppose the Amihud index of risk asset i is triangular fuzzy number $\hat{l}_i=\left[a,b,c\right]$, then the fuzzy liquidity of portfolio is $\sum_{i=1}^{n}\hat{l}_i x_i$, which reflects the liquidity of portfolio. By definition, the average likelihood of portfolio liquidity is [13]:

$$E\left(\hat{l}\left(x\right)\right)=E\left(\sum_{i=1}^{n}\hat{l}_i x_i\right) \tag{9}$$

The fuzzy Amihud index level $\hat{l}_0$ of a given investment sum can be obtained from the comparison rule:

$$\sum_{i=1}^{n}\hat{l}_i x_i \le \hat{l}_0 \cdots \Leftrightarrow \cdots E\left(\sum_{i=1}^{n}\hat{l}_i x_i\right)\le E\left(\hat{l}_0\right) \tag{10}$$

According to the linear programming game model theory of fuzzy decision-making [14], we refer to and modify the model used by Wu et al. [15]. At this point, we introduce two objective functions, one is to maximize the return of investors, the other is to minimize the risk. Therefore, we get the linear programming model for solving the optimal solution as follows:

$$Max \cdot \frac{\sum_{i=1}^{n+1} R_i x_i-\sum_{i=1}^{n} k_i\left|x_i\right|-M_0}{N_1-N_0}$$

$$Min \cdot \frac{M_1-\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j x_i x_j C\,\mathrm{ov}\left(R_i R_j\right)}{M_1-M_0}$$

$$s.t\begin{cases} E\left(\sum_{i=1}^{n}\hat{l}_i x_i\right)\ge E\left(\hat{l}_0\right) \\ p_0 \le \sum_{i=1}^{n} x_i \le p_0+Q_0 \\ x_i \ge 0, i=1,2,\ldots,n+1 \\ \quad 0\le \mu \le 1 \end{cases} \tag{11}$$

$$s.t\begin{cases} \dfrac{\sum_{i=1}^{n} R_i x_i - \sum_{i=1}^{n} k_i \left|x_i\right| - N_0}{N_1 - N_0} \geq \mu \\ \dfrac{M_1 - \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_i \sigma_j x_i x_j C\,\mathrm{ov}\left(R_i, R_j\right)}{M_1 - M_0} \geq \mu \\ p_0 \leq \sum_{i=1}^{n} x_i \leq p_0 + Q_0 \\ E\left(\sum_{i=1}^{n} \hat{l}_i x_i\right) \leq E\left(\hat{l}_0\right) \\ x_i \geq 0, i = 1, 2, \ldots, n+1 \\ 0 \leq \mu \leq 1 \end{cases} \tag{12}$$

Among them, x_i is the allocation weight of the i asset, p_0 and $p_0 + Q_0$ are the upper and lower limits of the proportion of funds available for investment determined by investors at a certain time according to the performance of the stock market. R_i is the current actual rate of return of asset i. σ_i^2 is the actual variance of the yield of the i asset.

It is assumed that the transaction cost is a V-shaped function. According to the transaction fee rules of 2021 in China, the stamp fee is 0.001 of the total amounts, the commission fee fluctuates between 0.001-0.003, and the transfer fee is 0.002. This paper mainly explains the practicability and effectiveness of the model, so it reasonably simplifies the calculation and takes the proportion of the whole transaction fee k = 0.005.

The first equation indicates that the satisfaction of the decision-maker does not exceed the satisfaction brought by the known portfolio. The second equation represents the degree of satisfaction that the risk can achieve within a subjectively given range; The third equation shows that the total investment proportion of risk asset portfolio is between its own capital and the maximum financing proportion; The fourth equation means that stocks cannot be sold short; The fifth equation is derived from the nature of membership.

Amihud index is a negative index. Investors have certain requirements for investors' liquidity, that is, it cannot be too low, otherwise the portfolio will be unable to trade and continue to lose money when the price fluctuates violently. The corresponding fuzzy Amihud index cannot be too high and needs to be less than a certain limit.

N_0, M_0, N_1, M_1, generally, it is given by experts or investors according to their own wishes and in combination with the actual situation. In this paper, based on the historical rate of return and the calculated semi deviation risk value, and referring to the empirical value given by Chen and other scholars [16], and making some modifications, take $M_0 = \frac{0.48}{\sqrt{12}} = 0.139, M_1 = \frac{0.08}{\sqrt{12}} = 0.023$ (assuming that the rates of return between months are independent and have the same distribution). In addition, the value of is determined by the opportunity cost of funds held by investors, which is equal to the one-month yield of government bonds, namely $N_0 = 0.012$. We also assume that the maximum expectation of investors' monthly return is $N_1 = 0.3$. When it is greater than this value, investors are unwilling to take more risks for excess return and will choose to cash in the return.

3. Experiment

3.1. Data notation

The experimental data source is CSMAR database. This paper collects the monthly rate of return (considering cash dividends), Amihud index and book to market ratio (BM) of 50 stocks from March 2020 to February 2021. We eliminate five stocks for the reason of data completeness. Then, we calculate the monthly variance of each stock during this period according to the monthly return on investment.

3.2. Process description

The experiment selects the first 20 stocks, and the labels of the 20 stocks are consistent with the sequence of their stock codes, as shown in the table 1.

Table 1: 20 stock s' labels and names

Label	Stock name	Label	Stock name
0	Ping A Bank	**10**	Merchants Shekou
1	Shenzhou High-speed Railway	**11**	Yifan Medicine
2	Yunnan Medicine	**12**	Jingxin Pharmaceutical
3	Luzhou Old Cellar	**13**	Topbond Shares
4	Gree Electric Appliance	**14**	Western Materials
5	BOE	**15**	IFLYTEK
6	Xinhua Pharmaceutical	**16**	Jiuyang Co., Ltd
7	China Color Co., Ltd	**17**	Shengnong Development
8	Wuliangye	**18**	Yanghe Co., Ltd
9	Yunnan Copper	**19**	Nanshan Holdings

Ward algorithm is used to extract the clustering features of these stocks, and the results are shown in Figure 1. We take the depth h = 4.3. At this depth, the number of clusters k = 4, so we get four types of stocks with different characteristics.

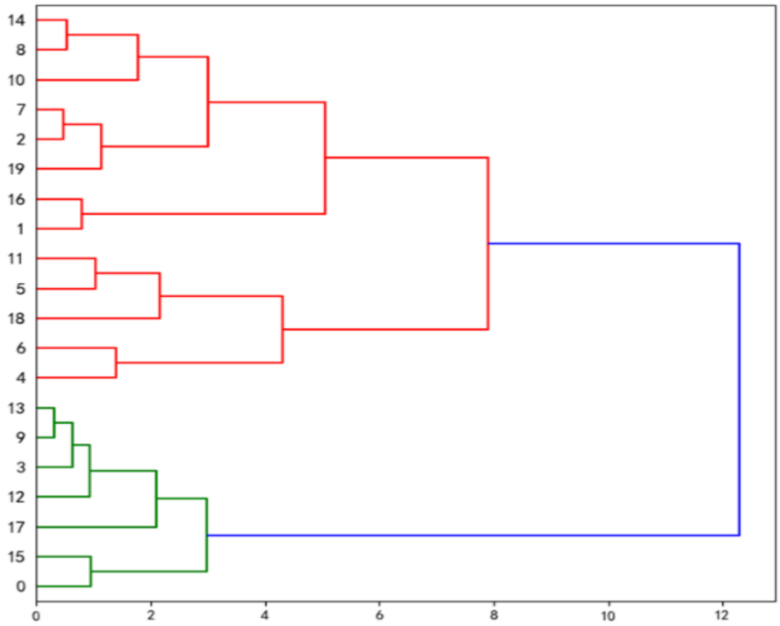

Figure 1: Clustering results of 20 stocks

Class I (Representative: Wuliangye and Nanshan Holdings)

Western materials, Wuliangye, China Merchants Shekou, China National Color Corporation, Yunnan Medicine, Nanshan Holdings

Class II (Representative: Shenzhou High-speed Railway and Jiuyang Shares)

Shenzhou High-speed Railway and Jiuyang Co., Ltd

Class III (Representatives: Yanghe Co., Ltd. and Xinhua Pharmaceutical)

Yifan Pharmaceutical, BOE, Yanghe Co., Ltd., Xinhua Pharmaceutical, Gree Electric Appliance

Class IV (Representatives: iFLYTEK, Luzhou Old Cellar)

Ping A Bank, iFLYTEK, Shengnong Development, Jingxin Pharmaceutical, Luzhou Old Cellar, Yunnan Copper and Tuobang Co., Ltd

Among them, it is found that four types of stocks formed by clustering have certain characteristics, as shown in Table 2.

Table 2: The features of four portfolios

	BM	Profit	Risk	Liquidity
1	High value investment	Average performance	Average risk	Low
2	Secondary	Poor performance	Average risk	High
3	Low	Average performance	High risk	Secondary
4	Secondary	Strong performance	High risk	Secondary

In the process of hierarchical clustering, some classes with poor performance or different investment style can be eliminated according to personal preferences. As a result, they can get a better portfolio which is more in line with their own investment style.

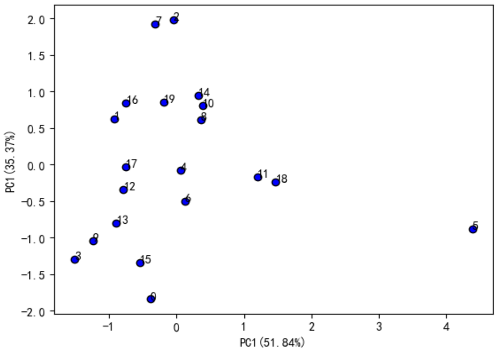

Figure 2: PCA analysis of 20 stocks

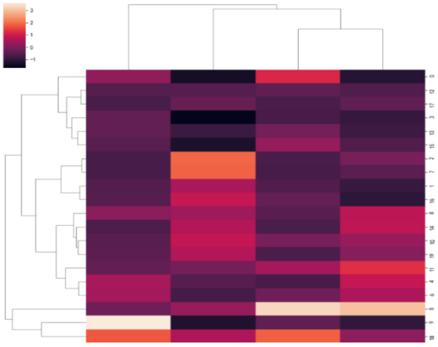

Figure 3: Heatmap Clustering of Ward's Algorithm

This paper also uses the principal component analysis method to analyze these 20 stocks, in which the first two principal component factors are selected, and the explanation degree is more than 85%, and the result is shown in Figure 2. It is very similar to the results of hierarchical clustering. For example, Stock 2 and Stock 7 are the closest, and they are also the first to be clustered together in hierarchical clustering. Besides, it is found that although the three stocks, 5, 8 and 11 are forcibly divided into one class when h = 4.5, Stock 5 is far from the other two stocks. Therefore, we can eliminate stock 5 or take it as a category alone.

The advantage of hierarchical clustering is that it shows the clustering process very intuitively. The clustering effect can be observed more intuitively. Understanding the structure of these data at the vertical level will help us understand its significance. In addition, hierarchical clustering also helps us adjust the clustering.

Similarly, similar conclusions can be obtained by using Ward algorithm for thermodynamic map clustering in this paper, which can be seen in Figure 3. In addition, in this example, we can also find that it is clustered first according to the rate of return and variance, then according to the liquidity, and finally according to the book value ratio. Thermodynamic diagram clustering clearly shows the structure of clustering.

3.3. Further analysis

This paper selects a larger data set for experiment. In this paper, 35, 45 stocks are also clustered respectively. The results of 45 stocks obtained are shown in Figure 4.

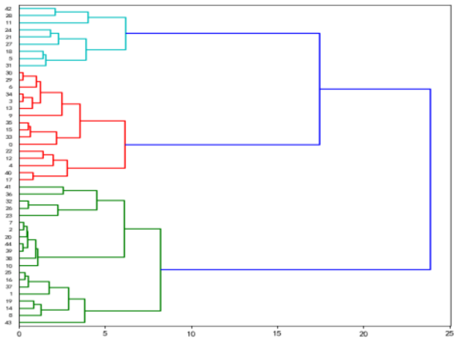

Figure 4: Clustering results of 45 stocks

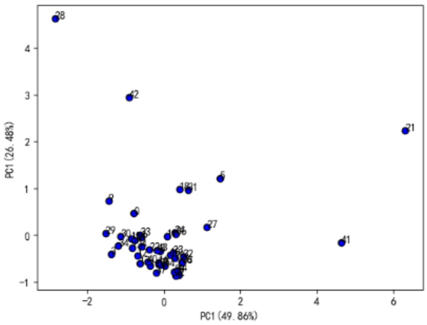

Figure 5: PCA analysis of 45 stocks

After the experiment of 35 and 45 stocks, it is found that, with the increasing number of sample data,

$\boldsymbol{\mu}$ becomes larger and larger. In the experiment with 35 stocks, it reaches 0.7396. In the experiment with 45 stocks, it reaches 0.8368.

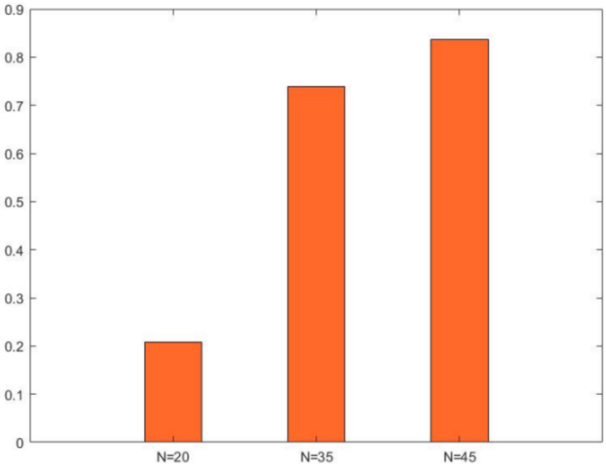

Figure 6: μ score comparison

4. Conclusion

In exploring the weight of portfolio allocation, this paper proposes a method of combining hierarchical clustering and fuzzy theory. And the $\boldsymbol{\mu}$ scores change according to the sample size studied. The experimental results show that when the sample size of the model is 20, 35 and 45 respectively, the combination weight score of hierarchical clustering gradually increases, indicating that investors are more and more satisfied, which shows the theoretical basis of large sample stock selection referred in many papers.

Compared with the original research, this paper combines hierarchical clustering and fuzzy theory, and takes double objectives as the objective functions. This not only breaks through the limitation of sample selection, but also realizes "personal stock selection"; At the same time, it is closer to the reality. The method given in this paper has certain reference.

References

[1] Markowitz H M. Portfolio Selection, Journal of Finance [J], 1952, 7
[2] Wang Zhengfang, Zhao Wenming, Ni Dejuan Portfolio and fuzzy programming, practice and understanding of mathematics [J], 1999. 1
[3] Xie Wenjun, Zhang Peng Mean semi absolute deviation interval portfolio optimization with entropy constraint, investment angle [J], 2018.
[4] Raffinot, T. (2017) Hierarchical Clustering-based Asset Allocation. The Journal of Portfolio Management, 44, 89-99.
[5] H. Tanaka and P. Guo, Portfolio selection based on upper and lower exponential possibility distributions, European Journal of Operational Research, 1999. Vol.114, 115-126.
[6] Tanaka and P. Guo and I.B. Turksen, Portfolio selection based on fuzzy probabilities and possibility distributions, Fuzzy sets and Systems, 2000, Vol.111, 387-397.
[7] Carlsson, R. Fuller, on possibilistic mean value and variance of fuzzy numbers, Fuzzy Sets and Systems, 2001, Vol. 122, 315-326.
[8] C. Carlsson, R. Fuller, P. Majlender, A possibilistic approach to selecting portfolios with highest utility scores, Fuzzy Sets and Systems, 2002, Vol. 131, 13-21
[9] W. Fang, Fuzzy Portfolio Efficiency Evaluation and decision analysis considering investor psychology based on DEA method [D]. master's thesis of South China University of technology, 2020.
[10] Li Meijuan, Yi Sicheng, Qiu Qirong, Lin Qi Dynamic triangular fuzzy multi-attribute decision-making method based on TOPSIS, system science and mathematics [J], 2021.
[11] Financial leverage and systematic risk analysis: time series analysis published by Amihud in 1977. Joe H. Ward, Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association [J], 2012, 236-244

[12] Feng Xiaohui, fuzzy portfolio optimization model considering securities transaction cost and liquidity [D], master's thesis of Hebei University of technology, 2009.

[13] Li Rongjun Theory and application of fuzzy multi criteria decision making [M]. Beijing: Science and Technology Press, 2002.

[14] Wu Jingui, Li Rongjun Research and application of fuzzy decision portfolio model with financing, Industrial Engineering Journal [J], 2011.6.

[15] Chen Wei, Zhang Runtong, Yang Ling. A fuzzy decision method for portfolio selection under the condition of existing financing [J]. Journal of Beijing Jiaotong University: Social Science Edition, 2007, 6 (1): 67-70.