Research on Stock Allocation Scheme Based on Two-step Clustering Algorithm

Mengdan Yu
School of Economics and Management
Beijing Jiaotong University
Beijing, China

Tao Chen
School of Economics and Management
Beijing Jiaotong University
Beijing, China

Jiayi Yao
School of Economics and Management
Beijing Jiaotong University
Beijing, China

Abstract—With the development of economy, the stock market has become an important part of China’s capital market, and the investment of stocks has become an important means for people to allocate assets. Therefore, how to select stocks, balance risk and return, and rationally optimize the allocation of stocks are of great significance. This paper chooses the sample data of CSI 300 Index of stock which are more suitable for the analysis of value investment theory, and uses two-step clustering algorithm to subdivide the stocks, so as to provide the direction of decision-making for the allocation of stocks.

Keywords—stock allocation; clustering analysis; two-step clustering

I. INTRODUCTION

Stock allocation is to select and match one or a group of stocks according to their own risk tolerance and investment style to achieve the reasonable expectation of investment risk and return through the analysis of stocks and investment demand. However, the securities market itself is a very complicated system, which contains many factors such as high noise, non-linearity and investors' arbitrariness. These factors determine the complexity of stock price trend prediction. Therefore, stock allocation decision-making has always been one of the hot research fields in society and academia [1].

Cluster analysis is an important method of multivariate data analysis to study classification. With the rapid development and application of data technology, cluster analysis has a very wide range of application scenarios. Through the study of previous history, it is found that people mainly rely on experience and professional knowledge to carry out qualitative analysis of the problems studied. The results often have strong subjectivity, and cannot objectively reveal the differences in the intrinsic nature of things. Especially for multi-index and multi-variable classification problems, qualitative analysis is difficult to achieve accurate classification [2]. Cluster analysis is then produced.

II. CONSTRUCTION OF TWO-STEP CLUSTER ANALYSIS MODEL

In this paper, two-step clustering algorithm is used to construct the clustering model. The principle of two-step clustering algorithm is as follows:

It is supposed that there are N data objects in data set D. Each data object has D attributes, including D1 continuous attributes and D2 subtype attributes. Let \( X_n = (x_{n1}, \ldots, x_{nD1}, y_{n1}, \ldots, y_{nD2}) \), where \( x_{n} \) represents the value of the nth data object under the sth continuous variable, and \( y_{nt} \) represents the value of the nth data object under the t-th type attribute. It is known that the t-th subtype attribute has \( c_t \) kinds of possible values. \( C_j = \{C_{j1}, \ldots, C_{jt}\} \) denotes a cluster having a cluster number J of the data set D, where \( C_j \) denotes the jth cluster in the cluster \( C_j \).

Two-step clustering algorithm is divided into two stages [3] [4]:

- Pre-clustering stage. Using the idea of CF tree growth in the principle of hierarchical algorithm, this algorithm will pre-classify the data points in dense areas and generate many small clusters when generating CF tree. Firstly, data points in data set D are inserted into the CF feature tree one by one to make it grow; when the size of CF feature tree exceeds the predetermined size, the algorithm will first remove potential outliers in the current CF feature tree, increase the spatial threshold and thin the CF feature tree, and then insert the outliers of the volume of the CF feature tree into the CF feature tree.
After traversing all the data, the potential outliers that cannot be inserted into the CF feature tree are the real outliers. Then the clustering features of the final CF leaf element corresponding to the sub-cluster are output to the next stage of the algorithm.

- **Clustering stage.** Taking the result of pre-clustering stage - sub-clusters as the object, the clustering method is used to merge sub-clusters one by one until the expected number of clusters. The input of this stage is the sub-cluster of the leaf element items of the final feature tree output in the pre-clustering stage, set to \( C_1, \ldots, C_l \). In fact, it is not a sub-cluster containing specific data points, but the clustering characteristics of each sub-cluster. Therefore, the work of this stage is based on the input data to the sub-cluster \( C_1, \ldots, C_l \) to carry out two-degree clustering to achieve the expected cluster number clustering results.

The two-step clustering algorithm uses the idea of condensation to merge the nearest clusters recursively to achieve the above purpose. First, it is necessary to find the two sub-clusters closest to each other from the \( j \) sub-clusters \( C_1, \ldots, C_j \). Then the number of clusters in the new clustering becomes \( j_1 \) less than the original one; then it is necessary to merge the nearest pair of clusters in the remaining clusters, repeat this operation until all the sub-clusters merge into a large cluster and get a cluster with cluster number of 1, so the clustering is \( C_1, \ldots, C_l \); finally, according to the purpose of the study, generate the expected number of clusters from the \( J \) cluster.

### III. EMPIRICAL ANALYSIS

#### A. Data Acquisition and Processing

The indicators in this paper mainly select the company's financial indicators such as EPS, net asset value per share, operating indicators such as net margins, net asset turnover, fundamental indicators such as P/E ratio, total market value and other internal and external indicators as the main input variables; At the same time, the data of CSI 300 listed companies are selected as samples, and the difference of the rise and fall between CSI300 index stocks and individual stocks is taken as the target variable of the model. (As shown in Table I)

| Serial number | Factor | Type              | Serial number | Factor | Type              |
|---------------|--------|-------------------|---------------|--------|-------------------|
| 1             | Net asset value per share | Continuity | 14           | Net interest rate in sale | Continuity |
| 2             | EPS    | Continuity        | 15           | Total operating cost rate | Continuity |
| 3             | Pre-tax profit per share  | Continuity | 16           | Main Business Ratio | Continuity |
| 4             | Current ratio      | Continuity | 17           | Return on total assets | Continuity |
| 5             | Quick ratio       | Continuity | 18           | Net asset turnover | Continuity |
| 6             | Cash ratio        | Continuity | 19           | P / B ratio | Continuity |
| 7             | Percentage of long-term liabilities | Continuity | 20           | P / E ratio | Continuity |
| 8             | Net profit growth rate | Continuity | 21           | Receivable turnover | Continuity |
| 9             | Annual growth rate of net assets | Continuity | 22           | Asset-liability ratio | Continuity |
| 10            | Annual growth rate of EPS | Continuity | 23           | Total market value | Continuity |
| 11            | Annual Growth Rate of Operating Revenue | Continuity | 24           | Circulated market value | Continuity |
| 12            | Net margins       | Continuity | 25           | Business income | Continuity |
| 13            | ROE               | Continuity | 26           | Earnings performance | Continuity |

Financial indicators, operating indicators and other internal indicators are selected from the four-year data from 2014 to 2017; basic indicators select the data at the end of 2017 for follow-up analysis; at the same time, the target variables are listed companies' stock price rise and fall, and the data select 2017 stock price rise and fall. Some stock and index data are shown in "Table II".
In the selection of factors, this paper mainly chooses financial indicators, operational indicators and other related basic indicators that can reflect the intrinsic value of the company. There are 26 variables involved in the selected data, which belongs to high-dimensional data, so it is necessary to reduce the dimension of the data. In this paper, the principal component analysis method is used to reduce the dimension and reconstruct the factor [5]. In the process of establishing the model, the new factors include solvency factor, per share index factor, scale factor, growth factor, profitability factor, price-earnings ratio factor, leverage factor and net assets factor as input variables, and the earnings performance of listed companies is classified. The earnings performance of listed companies, a continuous numerical variable, is discretized. Based on the increase of the CSI 300 index in 2017, the earnings performance of listed companies whose earnings performance is greater than or equal to the CSI 300 index is recorded as Category A, and whose earnings performance is less than the average is recorded as Category B.

Some data processing results are shown in "Table III"; and partial continuous variable data are shown in "Table IV".

### Table II. Partial Stock and Index Data

| Stock code   | Securities name         | Net asset value per share | EPS | Pre-tax profit per share | Current ratio | Quick ratio | Cash ratio | Percentage of long-term liabilities |
|--------------|-------------------------|---------------------------|-----|--------------------------|---------------|-------------|------------|-------------------------------------|
| 600004.SH    | Baiyun Airport          | 8                         | 1.02| 1.37                     | 1.12          | 1.07        | 0.76       | 0.39                                |
| 600009.SH    | Shanghai Airport        | 11.18                     | 4.44| 1.84                     | 4.54          | 4.53        | 4.03       | 0.55                                |
| 600010.SH    | Baotou Steel Group      | 1.29                      | 0.02| 0.03                     | 0.46          | 0.2         | 0.1        | 0.14                                |
| 600011.SH    | Huanneng Power International | 5.01                     | 0.6 | 1.54                     | 0.3           | 0.23        | 0.08       | 0.41                                |
| 600018.SH    | SINOPEC                 | 2.65                      | 0.34| 0.53                     | 0.97          | 0.62        | 0.49       | 0.35                                |
| 600019.SH    | Baoshan Iron & Steel    | 7.13                      | 0.46| 0.68                     | 0.82          | 0.35        | 0.11       | 0.11                                |
| 600023.SH    | Zeng Neng electric power| 4.13                      | 0.46| 0.73                     | 1.4           | 1.17        | 0.83       | 0.57                                |
| 600025.SH    | Huanneng Hydropower     | 2.04                      | 0.15| 0.42                     | 0.12          | 0.11        | 0.03       | 0.68                                |
| 600027.SH    | Huanneng Power International | 4.11                     | 0.49| 1.37                     | 0.31          | 0.25        | 0.1        | 0.49                                |
| 600028.SH    | SINOPEC                 | 5.64                      | 0.37| 0.64                     | 0.77          | 0.4         | 0.21       | 0.18                                |

### Table III. Section of Stock Profit Performance

| Stock code   | Securities name         | Profit Performance Continuity | Profit Performance Classification |
|--------------|-------------------------|-------------------------------|---------------------------------|
| 600004.SH    | Baiyun Airport          | 4.33%                         | B                               |
| 600009.SH    | Shanghai Airport        | 69.72%                        | A                               |
| 600010.SH    | Baotou Steel Group      | -11.83%                       | B                               |
| 600011.SH    | Huanneng Power International | -12.48%                    | B                               |
| 600018.SH    | SINOPEC                 | 29.88%                        | A                               |
| 600019.SH    | Baoshan Iron & Steel    | 36.06%                        | A                               |
| 600023.SH    | Zeng Neng electric power| -1.84%                        | B                               |
| 600025.SH    | Huanneng Hydropower     | 69.87%                        | A                               |
| 600027.SH    | Huanneng Power International | -25.05%                    | B                               |
| 600028.SH    | SINOPEC                 | 13.31%                        | B                               |
| 593500       | CSI 300 Index           | 25%                           |                                 |

### Table IV. New Factor Data of Stocks

| Stock code   | Securities name         | Solvency factor | Factor of Per Share Index | Scale factor | Growth factor | Profitability factor | P/E factor | Leverage factor | Asset factor |
|--------------|-------------------------|-----------------|---------------------------|--------------|--------------|----------------------|------------|-----------------|-------------|
| 600004.SH    | Baiyun Airport          | 0.07            | 0.19                      | -0.4         | -0.06        | 0.38                 | -0.78      | -1.09           | 0.11        |
| 600009.SH    | Shanghai Airport        | 2.01            | 0.42                      | -0.03        | -0.23        | 0.61                 | -0.82      | -2.21           | -0.1        |
| 600010.SH    | Baotou Steel Group      | -0.71           | -0.2                      | 0.04         | 1.82         | -0.06                | 3.99       | -0.02           | -0.02       |
| 600011.SH    | Huanneng Power International | -0.82          | -0.05                     | 0.03         | -0.27        | -0.15                | -0.31      | -0.86           | 0.21        |
| 600018.SH    | SINOPEC                 | -0.14           | -0.42                     | 0.46         | -0.09        | 0.73                 | -0.46      | -0.84           | 0.22        |
| 600019.SH    | Baoshan Iron & Steel    | -0.57           | -0.13                     | 1            | 0.87         | -0.35                | 0.24       | -0.22           | -0.28       |
| 600023.SH    | Zeng Neng electric power| -0.01           | -0.28                     | 0.03         | -0.25        | 0.05                 | -0.52      | -1.67           | 0.18        |
| 600025.SH    | Huanneng Hydropower     | -0.86           | -0.34                     | -0.2         | 0.1          | 0.3                  | 0.37       | -2.04           | 0.31        |
| 600027.SH    | Huanneng Power International | -0.88         | -0.11                     | -0.29        | -0.24        | -0.19                | -0.3       | -1.16           | 0.27        |
| 600028.SH    | SINOPEC                 | -0.32           | =1.71                     | 7.35         | -0.32        | -0.11                | -0.4       | 0.75            | -0.12       |

### B. Constructing Two-step Clustering Model

Based on the principle of the two-step clustering algorithm mentioned above, a clustering model is constructed by SPSS Modeler using two-step clustering algorithm. The output of the model is shown in "Fig. 1":...
From the classification results, the model divides the data into three categories: clustering-1 accounting for 56.1%, clustering-2 accounting for 36.0%, and clustering-3 accounting for 7.9%. In terms of classification quality, SPSS output results show that the classification quality is good.

Observing the clustering chart (see "Fig. 2"), the profit performance of cluster-1 which is shown in the category distribution chart, indicates that the data of A accounts for 100%. The profit performance of Listed Companies in category A is better than that of CSI 300 index, so this category attribute is high-quality stocks. Cluster-2's earnings performance which is shown in the category distribution chart indicates that the proportion of B data is 100%. Category B is listed companies' earnings performance is inferior to that of the CSI 300 Index, and the attribute is poor stocks.
IV. RESULT ANALYSIS

A. Clustering Feature Description

The final goal of clustering analysis is to get the classes and to describe the characteristics of the classes clearly enough. Therefore, comparing cluster-1 with cluster-2, "Fig. 3" can be got:

Fig. 3. Category comparison chart.

According to the output of the two-step clustering model, it can be seen that the characteristics of clustering-1 and clustering-2 are obviously different. By comparing Cluster-1 and Cluster-2, it is found that the stocks with better profitability are listed as Cluster-1, which is relatively low in P/E factor and relatively high in scale factor, asset factor and index factor per share.

B. Stock Selection Strategy

Through the above analysis, the output results of two-step clustering show that stocks that have excellent growth ability, profitability, operation ability and debt paying ability for a long time and can bring greater investment returns to investors have better profitability, and the stocks that can bring higher returns have lower prices.

According to the analysis of two-step clustering results, the strategy of stock allocation provided by this method is to select stocks with low P/E ratio, large market value and circulating market value, and high return on net assets as far as possible in the process of stock selection. On this basis, stocks with good earnings per share and growth ability can also be selected.

V. CONCLUSION

This paper applies two-step clustering algorithm to the mining and research of stock allocation, which is a useful discussion. Through data mining, it can be found out that the rules of stock selection provide a reliable basis for the rational selection of stocks. However, limited to the relationship between lengths, this paper only makes a limited discussion on clustering analysis method in data mining. It needs to further use professional knowledge to explore the influencing factors of stock selection, so that the accuracy of prediction is higher, therefore, there is still a lot of room for improvement.

REFERENCES

[1] Liu Hui, Huang Jianshan. "Risk factors analysis of stock return rate in China A-share market: based on Fama-French three-factor model," Contemporary Economic Science, vol 4, pp. 27-31+125, 2013.
[2] Wang Miao, Chai Ruimin. "An improved decision tree classification attribute selection method," Computer Engineering and Application Computer Engineering and Application. pp.11-15, 2010.
[3] Chen Jinlin, Yang Lin. "Research on stock price status and change based on cluster analysis," Times Finance, pp. 349 + 345,2018.
[4] Yu Qingguo. "Application of cluster analysis in securities investment," Cooperative Economy and Technology, pp.48-50, 2017.
[5] Haorui, Zhang Yue. "A stock method based on factor analysis and clustering — Taking Shanghai and Shenzhen 300 index component stocks as an example.," Times Finance, pp.135, 2014.