Research on the investment information automation problem based on the clustering analysis method

Wei GAO[1], Ruzhen YAN[2]

[1]Business School, Sichuan Agricultural University, Chengdu, China

[2]Business School, Chengdu University of Technology, Chengdu, China

KEYWORD: Clustering analysis; Investment; China securities market

ABSTRACT: This paper develops a two-step clustering method, and uses the data of China securities market to analyze the factors influencing the portfolio selection, and constructing an investment portfolio based on this clustering method. The results show that the performance of this portfolio based on the two-step clustering method is better than the traditional mean-variance model. The results indicate that this method is very important for the portfolio selection.

INTRODUCTION

In recent years, the credit card business, as one of the important benefit for commercial banks, is developing rapidly. Although commercial banks owe to mass data in big data, most of commercial banks in China have not carried out the effective analysis of credit card customers, and the homogenization of competition means among the banks are outstanding. In order to obtain a higher efficiency, commercial banks must pay attention to customer relationship management and maximize customer value based on customer segmentation. The so-called customer seg-mentation, according to the differences among the characteristics of customer demand and purchasing behavior, buying habits, reputation and other aspects of the situation, is the classification process of dividing the customers into a plurality of groups of customers.

Meanwhile, with the sustained and rapid development of economy, the capital market obtained the rapid development. However, affected by macroeconomic policies and the investor factors, the speculative arbitrage activities are very frequent in China capital market. The investors urgently need the guidance of the relevant investment theory, so that can make a rational investment activity. Hence, it is very important to further study the modern portfolio theory for the investor. Cluster analysis has been widely applied in portfolio selection.

Portfolio selection is the study of how people should invest their wealth. It is a process of trading off risk and expected return to find the best portfolio of assets and liabilities. However, in the real situation, many characteristics of objects are not obviously. In order to solve this situation, this paper develops a cluster segment method based on the fuzzy t-Norm, considers the problem of how to construct a portfolio based on the bid-ask spread, depth, order imbalance, market value, profit margins, asset-liability ratio.

Portfolio selection problem has important academic value for social and economic development. Atkinson & Wilmott(1995) develops a portfolio selection problem with fixed and variable transactions costs. The solution of this problem is given by way of quasi-variational inequalities for the object function. Boone & Roehm (2002) analysis the portfolio selection problem of risky assets with a diagonal covariance matrix, upper bounds on all assets and transactions costs, and develop an algorithm for its solution. Meanwhile, they present the efficient portfolios under appropriate assumptions. Young (2006), Zeng et al. (2013) consider the portfolio selection problem with transaction costs and constraints on exposure to risk.

In this paper, we consider the problem of how to construct a portfolio based on the bid-ask spread, depth, order imbalance, market value, profit margins, asset-liability ratio, and develops a cluster segment method based on the fuzzy t- Norm.

CLUSTERING ANALYSIS

Cluster analysis has been widely applied in portfolio selection. Gao et al.(2011) proposed the FCEN and applied to customer segmentation. The main idea of FCEN is regarding Fuzzy Clustering Algorithm (FCM) as basic cluster, using a method which is similar to Bagging to produce multiple basic clusters. Then, t-norm method is used to integrate the multiple clustering results, getting the final clustering results. In the Generating stage of the basic cluster, because the space distribution of the customers may be very uneven, the clustering results may represent a local optimization instead of the global clustering results. In this paper, a technique which is similar to re-sampling technique based on bagging method is used to produce some training set (represent the number of basic cluster) from the original data set and cluster each training subset with FCM.

The processing of two-stage model is as follows:

Step1: Input data set $X_{n\times m}$, and encode the feature m which length by Binary code;

Step2: Using formula (1), to calculate the fitness value, then using the roulette selection method to select chromosomes which will inherit to the next generation;

Step3: Using single point crossover operator with cross probability P_c to operate the cross calculation;

Step4: Using basic bit mutation operation with P_m to change some gene values in individual coding series and form a new individual;

Step5: Repeat step2, 3, 4 until the number of valid features m' is selected;

Step6: Input data set $X'_{n\times m'}$, number of clustering centers k and re-sampling times M;

Step7: Through similar Bagging re-sampling technology to generate M training sets and applying FCM method to clustering to obtain k clustering center after stability. Redistributing all data sets X to obtain membership matrix $U^{(t)}(t=1,2,\cdots M)$;

Step8: Using formula (2) to calculate the samples i, j belong to the same cluster membership matrix $C_{ij}^{(t)}$ in each clustering result. Then using formula (3) to calculate average membership degree matrix $\bar{C}_{ij}$.

Step9: Applying FCM method again to similar matrix $\bar{C}$ to obtain the sample belonging to a class of membership matrix U^c;

Step10: According to formula (4) to the final clustering result of the data set.

PORTFOLIO SELECTION

Clustering analysis indicators

There is no single portfolio selection strategy that is best for all people. There are, however, some general principals, such as the principal of diversification, which apply to all risk-averse people. Markowitz(1952) develops the mean-variance model that is a quantitative trade-off between risk and expected return. In this model, there is not the any transaction cost. However, the transaction cost is essential in the real securities market. Hence, when the investor constructing the portfolio selection, they must consider the bid-ask spread, depth, order imbalance, market value, profit margins, and asset-liability ratio.

In this paper, the stock evaluation indexes include:

(1) Return on equity

Return on equity (ROE) measures the rate of return for shareholders' equity, and measures the efficiency of a firm at mean profits from each unit of shareholder equity, also known as net assets or assets minus liabilities. This index shows how well a company uses investments to generate earnings growth, and can be calculated as following:

$$ROE = \frac{NetR}{StockH} \tag{1}$$

(2) Quick ratio

In finance, the quick ratio mainly measures the ability of a company to use its near cash or quick assets to extinguish or retire its current liabilities immediately. A company with a quick ratio of less than 1 cannot currently fully pay back its current liabilities.

(3) Net profit grow rate

Net profit, also referred to as net income, is a measure of the profitability of a venture after accounting for all costs.

$$NPGR = \frac{NetP - LNetP}{LNetP} \tag{2}$$

where, $NetP$ refers to the net profit; $LNetP$ represents the net profit of last year; $NPGR$ refers to the net profit grow rate.

(4) Liquidity index

Liquidity refers to the ability to facilitate the purchase or sale of stocks without causing drastic change in the stock price. In a liquid market, selling quickly will not reduce the price much. In a relatively illiquid market, selling it quickly will require cutting its price by some amount. In this paper, turnover is used to measure the liquidity.

$$TR_t = \frac{1}{n}\sum_{i=1}^{n}\frac{volume_i}{dsmvosd_i} \times 100\% \tag{3}$$

(5)Order imbalance

A situation resulting from an excess of buy or sell orders for a specific security on a trading exchange, making it impossible to match the orders of buyer and seller. For securities that are overseen by a market maker or specialist, shares may be brought in from a specified reserve to add liquidity, temporarily clearing out excess orders from the inventory so that the trading in the security can resume at an orderly level. Extreme cases of order imbalance may cause suspension of trading until the imbalance is resolved.

$$Orderimb_t = \frac{1}{n}\sum_{i=1}^{n}\left|BV_i - SV_i\right| \tag{4}$$

Sample data

There were 30 stocks selected from Shanghai and Shenzhen stock market as samples for this study. Based on the samples, we estimate and analyze the fuzzy clustering ensemble.

Clustering analysis and result

The descriptive statistical results of this study are shown in table 1.

Table 1 Descriptive statistics

Variable	Mean	Meddle	Variance
ROE	9.1103	7.6784	7.0333
QR	1.3222	1.246	0.7355
NPGR	46.6468	18.2576	176.2601
TR	1.1208	1.0776	0.9708
Orderimb	12899.92	10565.11	9489.38

For the data of 30 stocks, the clustering analysis and result is shown using fuzzy clustering ensemble method by the MATLAB 7.0 software. The result is shown in Table 2.

Table 2 the result of fuzzy cluster

	I	II	III	IV	V
1	600005	601616	600030	600176	600219
2	600126	600290	600634	600016	600028
3	600011	600098	600000	600824	600104
4	600328	601028	601299	600064	600068
5	600612		600056	600261	600007
6	600192				600667
7					600113
8					600019
9					600230
10					600565

Comparative analysis of different portfolio

Portfolio selection theory is a theory of finance that attempts to maximize portfolio expected return for a given amount of portfolio risk, or equivalently minimize risk for a given level of expected return, by carefully choosing the proportions of various assets. This method is widely used in practice in the financial industry. This method is a mathematical formulation of the concept of diversification in investing, with the aim of selecting a collection of investment assets that has collectively lower risk than any individual asset. This is possible, intuitively speaking, because different types of assets often change in value in opposite ways. For example, to the extent prices in the stock market move differently from prices in the bond market, a collection of both types of assets can in theory face lower overall risk than either individually. But diversification lowers risk even if assets' returns are not negatively correlated—indeed, even if they are positively correlated. This method can be shown as:

$$\min_{\{w\}} \ \frac{1}{2} w^T V w$$

$$s.t. \ \ w^T e - w^T \beta = \mathrm{E}\left[\tilde{r}_p\right] \qquad (5)$$

$$w^T \mathrm{I} = 1$$

For the data from the China securities market, the optional portfolio is based on the traditional mean-variance method. Supposed the holding period is one season of the investor, the expected revenue can be shown in Table 3.

Table 3 The security price and invest ratio

	600005	600126	600011	600328	600612	600192
p0	2.20	5.06	3.88	6.86	7.03	23.32
P1	2.05	5.04	3.60	7.74	7.63	23.96
r	0.1182	0.0863	0.1514	0.1465	0.2593	0.2383

The expected return can be calculated by the formulation:

$$TR = \sum_{i=1}^{m} r_i \frac{P_{i1} - P_{i0}}{P_{i0}} \qquad (6)$$

The return is 2.26% in 2014 for the investor based on the traditional mean-variance method. The optional portfolio is P_2 based on the two-step clustering method. Supposed the holding period is one season of the investor, the expected revenue can be shown in Table 4:

Table 4 The security price based on the clustering method

	600000	600016	600030	600126	600192	601299
p0	9.43	7.72	12.75	3.88	6.86	4.92
P1	9.72	7.66	10.53	3.6	7.74	4.64
r	0.241	0.0792	0.1273	0.1183	0.2861	0.1481

The return is 0.44% in 2014 for the investor based on the two-step clustering method. Obviously, the performance of portfolio using fuzzy clustering method is better than the portfolio based on the traditional method.

CONCLUSIONS

Portfolio selection is the study of how people should invest their wealth. It is a process of trading off risk and expected return to find the best portfolio of assets and liabilities. Cluster analysis is the task of grouping a set of objects in such a way that objects in the cluster are more similar to each other than to those in other clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, portfolio selection.

This paper develops a cluster segment method based on the fuzzy t-Norm, considers the problem of how to construct a portfolio based on the bid-ask spread, depth, order imbalance, market value, profit margins, asset-liability ratio. The method from this paper provides guidance in portfolio selection for the investors. The performance of portfolio using fuzzy clustering method is better than the portfolio based on the traditional method.

ACKNOWLEDGEMENTS

This paper is supported by the Key Projects of Educational Commission of Sichuan Province of China (13ZA0140), and the Program of Undergraduate Thesis Foster of Sichuan Agriculture University (34009114).

REFERENCES

[1] Atkinson C & Wilmott P. 1995. Portfolio management with transaction costs: an asymptotic analysis of the morton and pliska model. *Mathematical Finance* 5(4): 357-367.

[2] Boone D. S. & Roehm M. 2002. Retail segmentation using artificial neural networks. *International Journal of Research in Marketing* 19: 287-301.

[3] Deng X. Y., Jin C. & Han Q. P. 2011. KSP: a hybrid clustering algorithm for customer segmentation in mobile E-commerce. *Journal of Management Science* 24: 54-61.

[4] Doyne J., Farmer L. & Gillemot F. 2004. What really causes large price changes?. *Quantitative Finance* 4(4): 383-397.

[5] Dudoit S. & Fridlyand J. 2003. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 19: 1090-1099.

[6] Fowlkes E. B., Gnanadesikan R. & Kettenring J. R. 1998. Variable selection in clustering, *Journal of Classification* 5: 205-228.

[7] Freytag P. V. 2001. Business to business market segmentation. *Industrial Marketing Management* 30: 473-486.

[8] Gao W., He C. Z. & Jiang X. Y. 2011. Customer segmentation study based on fuzzy clustering ensemble. J*ournal of Intelligence* 30: 125-129.

[9] Hasbrouck J. & Schwartz R. A. 1988. Liquidity and execution costs in equity markets. *The Journal of Portfolio Management* 14(3): 10-16.

[10] Kim J. 2003. Segmentation the market of West Australian senior tourist using artificial neural network. *Tourism Management* 24(1): 25-34.

[11] Konno H. & Yamazaki H. 1991. Mean-absolute deviation portfolio optimization model and its application to Tokyo stock market. *Management Science* 37: 519-531.
[12] Konno H. & Yamazaki H. 1991. Mean-absolute deviation portfolio optimization model and its application to Tokyo stock market. *Management Science* 37: 519-531.
[13] Korn R. 1997. Optimal portfolios: stochastic models for optimal investment and risk management in continuous time. *Singapore: World Scientific Publishing Co. Pte. Ltd*.
[14] Longstaff F. A. 2001. Optimal portfolio choice and the valuation of illiquid securities. *Review of Financial Studies* 14(2): 407-431.
[15] Markowitz H. 1952. Portfolio selection[J]. *Journal of Finance* 7(1): 77-91.
[16] Milligan G.W. & Cooper M .C. 1985. An examination of procedures for determining the number of cluster in a data set. *Psychometrika* 50: 159-179.
[17] Strehl A. & Ghosh J. 2002. Cluster Ensembles: a knowledge reuse framework for combination multiple partition. *Journal of Machine Learning Research* 3: 583-617.
[18] Weber P. & Rosenow B. 2005. Order book approach to price impact. *Quantitative Finance* 5(4): 357-364.
[19] Yang Y., Jin F. & Kamel M. 2008. Survey of clustering validity evaluation. *Application Research of Computer* 25: 1630-1633.
[20] Young M. R. 2006. A mini-max portfolio selection rule with linear programming solution. *Management Science* 44: 673-683.
[21] Zeng X. Q., Xu Q. & Zhang D. 2013. New multi-indicator customer segmentation method based on consuming data mining. *Application Research of Computers* 30: 2944-2947.
[22] Zhou Z. H. & Tang W. 2006. Cluster ensemble. *Knowledge-Based Systems* 19: 77-83.