Stock Portfolio Decision Based on Cluster Analysis and Principal Component Analysis

Yue Duan *

Beijing university of posts and telecommunications, Beijing, China.

duanyuezoe@gmail.com

*Corresponding author

Keywords: Portfolio theory, Diversification, Cluster analysis, Sharpe Ratio, Principal component analysis, Linear regression analysis.

Abstract. When determining the stock portfolio, it is necessary to consider factors such as the company's financial indicators, stock price trends, and macroeconomic conditions. In order to simplify the process, the clustering method is used to classify the company according to the financial indicator data, so as to select stocks when diversifying the investment, and then Sharp Ratio is calculated to conduct risk assessment on the selected targets. Finally, the principal component analysis method is used to simplify the obtained data, and then predict the future trend of the stock price by linear regression, and finally an ideal investment portfolio is determined.

Introduction

Among many financial products, stocks are the targets of many investors. The stock market is a very important part of the investment market. As we all know, options, bonds and funds are financial instruments based on stocks. Therefore, the fluctuation of stock prices and the overall situation of the stock market have been widely focused.

Before investing in the stock market, investors will observe the selected targets for a period of time with some methods. These methods are generally K-line observation or fundamental analysis. But the stock market is affected by many factors, and more accurate methods are needed to determine their investment targets in order to obtain more stable returns.

One of the important premises of portfolio theory is "diversified investment", which means we should diversify the capital into different types of companies to avoid risks and maximize the return. But sometimes it's hard to define "different type", so here we can select the financial indicators that reflect the company's situation and use clustering to classify them. Deng Xiuqin also used the clustering method in her article [3]. here we use a similar method to classify several selected targets, in Deng Xiuqin's article, only selected five indicators of profitability to reflect the company's type, and in this paper, we have made some adjustments to the financial indicators selected during clustering, taking into account the ability of a company's profit, debt repayment, and growth, therefore obtain more reasonable classification results.

After classifying it, analyze the previous data for different categories, and obtain the Sharpe Ratio [4] of each underlying stock to determine the size of the investment risk, which is helpful for us to choose stocks. It can help us choose a higher-yielding and more stable investment target within the acceptable risk range.

In the selected targets, the principal component analysis method is used to analyze the indicators reflecting the price changes, and several representative principal components are obtained. According to the results obtained in < Stock Price Forecast Based on Principal Component Analysis and Generalized Regression Neural Network> written by ZhuoXi Yu. [5] Based on Zhuoxi's paper, this article deletes the three indicators of earnings per share, return on equity, and net assets per share, because the operating conditions in the first part of the cluster analysis has been used the company's financial indicators, this is only part of a number of technical indicators to predict future stock price, and ultimately determine an appropriate investment portfolio based on the results.

Data Sources

In order to analyze the hot stocks, we selected the top 10 stocks of the big trade rankings on the day of Netease Finance as the sample when the data was selected. The 10 stocks were respectively Meidijituan (000333), Jiakaicheng (000918), Ningbohuaxiang (002048), Hengxingkeji (002132), Xibujianshe (002302), Lipenggufen (002374), Guochuangxingao (002377), Shandongjingmi (002384), Xinbangyaoye (002390), Huaweiwenhua (002502) .

The companies represented by these 10 stocks are involved in various sectors such as electrical appliance manufacturing, real estate, automobile manufacturing, metal products, construction, communications equipment, pharmaceutical manufacturing, and publishing media. The regions are also scattered in various provinces. The stocks can be regarded as hot stocks and spread across all sectors, and the results obtained should be reasonable and universal.

The collected information is determined by the stock's corresponding company's financial indicators for the third quarter of 2018 and the indicators which reflect price fluctuations in the past year. The source of the information is the statistics released by several financial websites such as Netease Finance, Sina Finance, Straight Flush, and Oriental Fortune. It is stored in Excel and is excluded from the abnormal value of the trading day.

Operations and Results

Stock Classification Based on Pedigree Cluster Analysis

There are many financial indicators that reflect the company's situation. Here, in order to comprehensively consider the capabilities of the company [6], we have selected the following nine indicators:

- Profitability
  - Rate of Return on Common Stockholders' Equity (ROE)
  - Earnings Per Share (EPS)
  - Operating Profit Ratio (OPR)
- Repayment Ability
  - Debt Asset Ratio (DAR)
  - Current Ratio (CR)
  - Quick Ratio (QR)
- Growth Ability
  - Main Income Growth Rate (MIGR)
  - Net Profit Growth Rate (NPGR)
  - Total Assets Growth Rate (TAGR)

The data of the indicators of the ten companies are summarized as follows:

Table 1

	ROE	EPS	OPR	DAR	CR	QR	MIGR	NPGR	TAGR
Meidijituan	21.60	2.72	26.67	64.16	1.42	1.23	10.06	18.56	5.97
Jiakaicheng	-11.93	-0.30	-2.38	81.21	1.20	0.11	-18.37	0	-25.28
Ningbohuaxian	6.31	0.8	18.78	40.55	1.51	1.04	-0.36	-18.98	11.89
Hengxingkeji	-0.63	-0.01	12.89	48.08	1.19	0.96	5.54	-128.18	-3.25
Xibujianshe	3.46	0.18	8.39	67.71	1.41	1.39	29.69	104.97	24.67
Lipenggufen	0.72	0.03	12.67	47.66	1.10	0.48	-24.66	-79.43	9.35
Guochuangxingao	4.41	0.24	15.93	22.45	1.63	1.45	181.04	450.16	-2.49
Dongshanjingmi	8.15	0.42	15.48	73.28	0.92	0.69	28.49	81.35	60.24
Xinbangzhiyao	4.04	0.16	21.12	45.80	1.26	1.07	14.83	9.23	6.48
Huaweiwenhua	0.97	0.04	79.23	10.04	3.71	2.24	-79.36	-87.01	6.04

After putting this data into SAS system, we use the pedigree cluster method to classify them and got the following result:

Table 2

Obs	company	CLUSTER	CLUSNAME
1	Ninbohuaxiang	1	CL5
2	Xinbangzhiyao	1	CL5
3	Henxingjituan	1	CL5
4	Lipenggufen	1	CL5
5	Xibujianshe	1	CL5
6	Dongshanjingmi	1	CL5
7	Jiakaicheng	2	Jiakaicheng
8	Meidijituan	3	Meidijituan
9	Guochuangxingao	4	Guochuangxingao
10	Huaweiwenhua	5	Huaweiwenhua

Here we divide the 10 targets into five categories according to the distance among them. The four samples of Jiakaicheng, Meidijituan, Guochuangxingao, Huaweiwenhua, indicating that they are far from other six samples, so they can't be divided into the same category. The rest samples are close to each other, so they are classified into one category.

Checking the nine indicators that reflect the company's status, it is not difficult to find that the six companies that are classified as same category are all with medium profitability and weak repayment ability but a good growth ability, which means the result obtained by cluster analysis is reasonable.

According to the portfolio theory, the development status of six companies in the same category is similar, so it should not be invested at the same time in order to avoid risks. When determining the investment portfolio, we can first select the four stocks which are not in the same category with any others, and then determine the fifth stock in the portfolio through subsequent calculation and analysis.

Sharpe Ratio of the Candidate Stocks

According to the above analysis, we have selected 4 targets, and we still need to select one of the 6 stocks to join the investment portfolio.

Here we downloaded the price limit data in the past year of each stocks, and the interest rate of risk-free investment, the abnormal data of non-trading days is changed to 0, and then analyzed.

We need to input the data into the SAS system, and then use the SAS function univariate to analyze the rise and fall of each stock, then get the expected value and standard deviation of each sample. The expected value is subtracted from the risk-free daily rate of return and divided by the standard deviation of the target. The results are as follows:

Table 3

	Expected value	Standard deviation	Sharpe Ratio
Ningbohuaxiang	-0.3168	2.4533	-0.1292
Hengxingkeji	-0.2699	1.8896	-0.1429
Xibujianshe	-0.2205	2.9058	-0.0759
Lipenggufen	-0.2022	2.8034	-0.0722
Dongshanjingmi	-0.1501	3.1785	-0.0473
Xinbangzhiyao	-0.2549	2.3699	-0.1064

It's obvious that the expected value of the six stocks is negative in 2018, which means that the situation of the six stocks is not very optimistic, but considering the overall downturn of the stock market in 2018, this phenomenon can be explained. These stocks may still have some investment value.

Therefore, when choosing, for ordinary investors of non-professional institutions, it is necessary to comprehensively consider the two aspects of risk and return, and select "high cost performance" investment, that is to say: the game of positive return should be less volatile. This kind of income is relatively stable and the risk is small. For stocks with negative returns, if the choice is fluctuating, the profit may be higher, but the stock with higher risk relative to the high return. In short, you should choose the stocks with higher Sharpe Ratio.

Based on the above analysis, among the six stocks, Dongshanjingmi, Lipenggufen, Xibujianshe are more stable for holding, but the specific choice of which stock to be in the portfolio needs further analysis.

Evaluation of Stocks Based on Principal Component Analysis and Linear Regression

There are many factors that can reflect the fluctuation of stock. When conducting stock price forecasting, it's necessary to select some of the more important parts for analysis. However, the result of manual selection isn't always correct, it may be partially redundant or insufficient, so it's hard to obtain accurate information of forecasting.

In order to further simplify the data that needs to be analyzed on the basis of fully reflecting the stock price information, this paper uses the principal component analysis method to deal with the data of opening price, closing price, highest price, lowest price and so on (Totally six factors). The closing price is regarded as the indicator of stock price forecast.

Here we take the stock Xibujianshe as an example to calculate. After inputting the data into SAS, use the function corr and princomp. In the corr results, we can get that there is a strong correlation between the original variables, which means that we have a lot of redundancy in the selected indicators, it is necessary to carry out principal component analysis.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.67692365	3.39208210	0.7795	0.7795
2	1.28484155	1.26821519	0.2141	0.9936
3	0.01662636	0.00438600	0.0028	0.9964
4	0.01224036	0.00530175	0.0020	0.9984
5	0.00693861	0.00450914	0.0012	0.9996
6	0.00242947		0.0004	1.0000

Fig.1

Combining the results of princomp in Fig.1, it's obvious that the cumulative contribution rate of the first two principal components has reached 99.36%, which can basically reflect all the information. Only two principal components can reflect all the information of six indicators, which greatly reduces the analysis work and improves the analysis efficiency while ensuring the accuracy.

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
x1	0.447843	-.203902	-.662914	0.221920	0.014585	0.518600
x2	0.452239	-.175618	-.077944	0.128097	-.621511	-.596555
x3	0.440062	-.264409	0.049659	-.099047	0.751501	-.399253
x4	0.440789	-.253094	0.701173	-.099075	-.160553	0.463046
x5	0.290567	0.683119	0.190727	0.627958	0.134500	-.011033
x6	0.350054	0.571806	-.154866	-.721378	-.069927	0.035225

Fig.2

We can get the function of the first two principal components Y_1, Y_2, which are:

$$Y_1=0.447843x_1+0.452239x_2+0.440062x_3+0.440789x_4+0.290567x_5+0.350054x_6 \quad (1)$$

$$Y_2=-0.203902x_1-0.175618x_2-0.264409x_3-0.253094x_4+0.683119x_5+0.571806x_6 \quad (2)$$

At this time, when we conduct stock evaluation, we do not need to carry out regression analysis on the six indicators, but directly use the two principal components obtained for analysis. When predicting the price, we use the next day's stock closing price as the dependent variable, and the two principal components corresponding to the data of the day as the independent variable.

After linear regression, the following results are obtained:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	603.72016	301.86008	210.23	<.0001
Error	237	340.29685	1.43585		
Corrected Total	239	944.01701			

Root MSE	1.19827	R-Square	0.6395
Dependent Mean	13.73062	Adj R-Sq	0.6365
Coeff Var	8.72699		

Fig.3

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.35088	0.14871	89.78	<.0001
x1	1	0.00000423	2.984952E-7	14.18	<.0001
x2	1	-0.00000253	1.790608E-7	-14.12	<.0001

Fig.4

In the results of parameter estimates, we get the equation:

$$Y = 0.00000423x_1 - 0.00000253x_2 \quad (3)$$

From the test of the P value and the data of the R square, it is not difficult to see that there is a significant relationship between Y and the two indicators from the previous data, so this function can be used to evaluate the potential value of each stock.

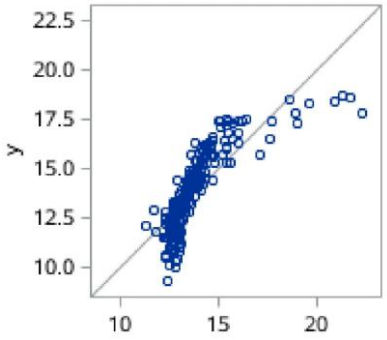

Fig.5

According to our stock price forecasting function, we can use some of today's technical indicators to predict the stock price of tomorrow, and judge whether this is a stock that can bring short-term profits. The above-mentioned method of principal component analysis and re-linear regression is applied to the remaining two stocks in turn, and the predicted values for tomorrow are calculated to determine which stock is more suitable for holding.

Conclusion

According to the research in this paper, it is found that when determining the portfolio, the cluster analysis is used to classify the companies to be selected, the companies that do not belong to the same category are selected for investment, and the Sharpe Ratio is used to compare the investment risks of the candidate stocks. Finally, the principal component analysis is used. And linear regression to find the stock price forecast equation of each stock, in turn to predict the stock price fluctuations in the next few days, and finally determine their own portfolio.

The whole process is effective and easy to achieve, in the choice of portfolio, you can base on the provisions provided in this article and the use the K-line analysis as well.

However, in the research of this paper, there are mainly the following problems, and further research is needed: The macroeconomic situation and policy changes in the company's location and industry are affected, but because these factors are difficult to measure with a certain indicator, this article did not consider these aspects in the analysis. This is not accurate enough. So when solving this random, non-linear problem, more research and learning is needed here.

References

[1] Markowitz, H.M, "Portfolio Selection", The Journal of Finance, 7(1): 77–91, doi: 10.2307/2975974. JSTOR 2975974. (1952)

[2] Markowitz, Harry, Mean-Variance Analysis in Portfolio Choice and Capital Markets. Wiley, ISBN 978-1-883-24975-5. (2000)

[3] Xiuqin Deng, Application of Cluster analysis in stock market board analysis, doi:10.13860/j.cnki.sltj.1999.05.001. (1998)

[4] Sharpe, William F, "The Sharpe Ratio". The Journal of Portfolio Management, 21(1): 49–58. doi:10.3905/jpm.1994.409501, Retrieved June 12, 2012.

[5] ZhuoXi Yu, Stock Price Forecast Based on Principal Component Analysis and Generalized Regression Neural Network, doi:10.13546/j.cnki.tjyjc.2018.18.039. (2018)

[6] Yang qi, Correlation Analysis of Stock Price of China's Financial Institutions and Main Financial Indicators, 2018, F830.42;F832.51.