Selecting Valuable Stock Using Genetic Algorithm

Chengxiong Zhou[1], Lean Yu[1,2], Tao Huang[3], Shouyang Wang[1], and Kin Keung Lai[2]

[1] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China
zcx1975@sina.com, {yulean, sywang}@amss.ac.cn
[2] Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
{msyulean, mskklai}@cityu.edu.hk
[3] School of Public Policy & Management, Tsinghua University, Beijing, 100084, China
ht01@mails.tsinghua.edu.cn

Abstract. In this study, we utilize the genetic algorithm (GA) to select high quality stocks with investment value. Given the fundamental financial and price information of stocks trading, we attempt to use GA to identify stocks that are likely to outperform the market by having excess returns. To evaluate the efficiency of the GA for stock selection, the return of equally weighted portfolio formed by the stocks selected by GA is used as evaluation criterion. Experiment results reveal that the proposed GA for stock selection provides a very flexible and useful tool to assist the investors in selecting valuable stocks.

1 Introduction

In the stock market, investors are often faced with a large number of stocks. A crucial work of their investment decision process is the selection of stocks. From a data-mining perspective, the problem of stock selection is to identify good quality stocks that are potential to outperform the market by having excess return in the future. Given the fundamental accounting and price information of stock trading, it is a prediction problem that involves discovering useful patterns or relationship in the data, and applying that information to identify whether a stock is good quality.

Obviously, it is not an easy task for many investors when they faced with enormous amount of stocks in the market. With focus on the business computing, applying artificial intelligence to portfolio selection and optimization is one way to meet the challenge. Some research has presented to solve asset selection problem. Levin [1] applied artificial neural network to select valuable stocks. Chu [2] used fuzzy multiple attribute decision analysis to select stocks for portfolio. Similarly, Zargham [3] used a fuzzy rule-based system to evaluate the listed stocks and realize stock selection. Recently, Fan [4] utilized support vector machine to train universal feedforward neural networks to perform stock selection.

However, these approaches have some drawbacks in solving the stock selection problem. For example, fuzzy approach [2-3] usually lacks learning ability, while neural network approach [1, 4] has overfitting problem and is often easy to trap into local minima. In order to overcome these shortcomings, GA is used to perform this task. Some related typical literature can be referred to [5-7] for more details.

T.-D. Wang et al. (Eds.): SEAL 2006, LNCS 4247, pp. 688 – 694, 2006.

The main aim of this study is to select some valuable stocks using GA and to test the efficiency of the GA for stock selection. The rest of the study is organized as follows. Section 2 describes the selection process based on the genetic algorithm in detail. Section 3 presents a simulation experiment. And Section 4 concludes.

2 GA-Based Stock Selection Process

Generally, GA imitates the natural selection process in biological evolution with selection, crossover and mutation, and the sequence of the different operations of a genetic algorithm is shown in the left part of Fig. 1. That is, GA is procedures modeled after genetics and evolution. Genetics provide the chromosomal representation to encode the solution space of the problem while evolutionary procedures are designed to efficiently search for attractive solutions to large and complex problem. Usually, GA is based on the survival-of-the-fittest fashion by gradually manipulating the potential problem solutions to obtain the more superior solutions in population. Optimization is performed in the representation rather than in the problem space directly. To date, GA has become a popular optimization method as they often succeed in finding the best optimum by global search in contrast to most common optimization algorithms. Interested readers can be referred to [8-9] for more details.

The aim of this study is to identify the quality of each stock using GA so that investors can choose some good ones for investment. Here we use stock ranking to determine the quality of stock. The stocks with a high rank are regarded as good quality stock. In this study, some financial indicators of the listed companies are employed to determine and identify the quality of each stock. That is, the financial indicators of the companies are used as input variables while a score is given to rate the stocks. The output variable is stock ranking. Throughout the study, four important financial indicators, return on capital employed (ROCE), price/earnings ratio (P/E Ratio), earning per share (EPS) and liquidity ratio are utilized in this study.

ROCE is an indicator of a company's profitability related to the total financing, which is calculated as

$$\text{ROCE} = (\text{Profit})/(\text{Shareholder's equity}) \times 100\% \tag{1}$$

The higher the indicator (ROCE), the better is the company's performance in terms of how efficient the company utilizes shareholder's capital to produce revenue.

P/E Ratio measures the multiple of earnings per share at which the stock is traded on the stock exchange. The higher the ratio, the stronger is the company's earning power. The calculation of this ratio is computed by

$$\text{P/E ratio} = (\text{stock price})/(\text{earnings per share}) \times 100\% \tag{2}$$

EPS is a performance indicator that expresses a company's net income in relation to the number of ordinary shares issued. Generally, the higher the indicator, the better is the company's investment value. The calculation of the indicator can be represented as

$$\text{Earnings per share} = (\text{Net income})/(\text{The number of ordinary shares}) \tag{3}$$

Liquidity ratio measures the extent to which a company can quickly liquidate assets to cover short-term liabilities. It is calculated as follows:

$$\text{Liquidity Ratio} = (\text{Current Assets})/(\text{Current Liabilities}) \times 100\% \quad (4)$$

If the liquidity ratio is too high, company performance is not good due to too much cash or stock on hand. When the ratio is too low, the company does not have sufficient cash to settle short-term debt.

When the input variables are determined, we can use GA to distinguish and identify the quality of each stock, as illustrated in Fig. 1. The detailed procedure is illustrated as follows.

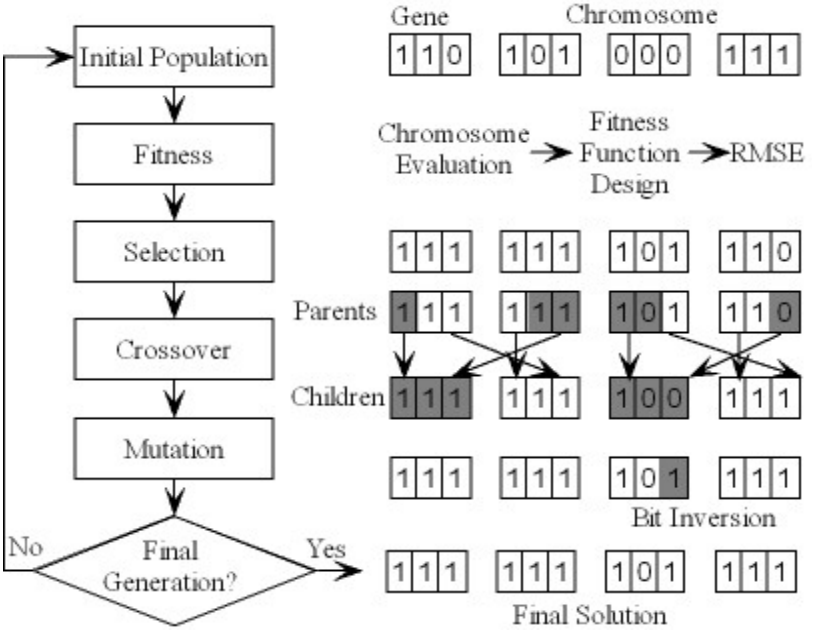

Fig. 1. Stock selection with genetic algorithm

First of all, a population, which consists of a given number of chromosomes, is initially created by randomly assigning "1" and "0" to all genes. In the case of stock ranking, a gene contains only a single bit string for the status of input variable. The top right part of Figure 1 shows a population with four chromosomes, each chromosome includes different genes. In this study, the initial population of the GA is generated by encoding four input variables. For the testing case of ROCE, we design 8 statuses representing different qualities in terms of different interval, varying from 0 (Extremely poor) to 7 (very good). An example of encoding ROCE is shown in Table 1. Other input variables are encoded by the same principle. That is, the binary string of a gene consists of three single bits, as illustrated by Fig. 1.

Table 1. An example of encoding ROCE

ROCE value	Status	Encoding
(-∞, -30%]	0	000
(-30%, -20%]	1	001
(-20%,-10%]	2	010
(-10%,0%]	3	011
(0%, 10%]	4	100
(10%, 20%]	5	101
(20%, 30%]	6	110
(30%,+∞)	7	111

Note that 3-digit encoding is used for simplicity in this study. Of course, 4-digit encoding is also adopted, but the computations will be rather complexity.

The subsequent work is to evaluate the chromosomes generated by previous operation by a so-called fitness function, while the design of the fitness function is a crucial point in using GA, which determines what a GA should optimize. Since the output is some estimated stock ranking of designated testing companies, some actual stock ranking should be defined in advance for designing fitness function. Here we use annual price return (APR) to rank the listed stock and the APR is represented as

$$APR_n = \frac{ASP_n - ASP_{n-1}}{ASP_{n-1}} \tag{5}$$

where APR_n is the annual price return for year n, ASP_n is the annual stock price for year n. Usually, the stocks with a high annual price return are regarded as good stocks. With the value of APR evaluated for each of the N trading stocks, they will be assigned for a ranking r ranged from 1 and N, where 1 is the highest value of the APR while N is the lowest. For convenience of comparison, the stock's rank r should be mapped linearly into stock ranking ranged from 0 to 7 with the following equation:

$$R_{actual} = 7 \times \frac{N - r}{N - 1} \tag{6}$$

Thus, the fitness function can be designed to minimize the root mean square error (RMSE) of the difference between the financial indicator derived ranking and the next year's actual ranking of all the listed companies for a particular chromosome, representing by

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(R_{derived} - R_{actual}\right)^2} \tag{7}$$

After evolving the fitness of the population, the best chromosomes with the highest fitness value are selected by means of the roulette wheel. Thereby, the chromosomes are allocated space on a roulette wheel proportional to their fitness and thus the fittest chromosomes are more likely selected. In the following crossover step, offspring chromosomes are created by some crossover techniques. A so-called one-point crossover technique is employed, which randomly selects a crossover point within the

chromosome. Then two parent chromosomes are interchanged at this point to produce two new offspring. After that, the chromosomes are mutated with a probability of 0.005 per gene by randomly changing genes from "0" to "1" and vice versa. The mutation prevents the GA from converging too quickly in a small area of the search space. Finally, the final generation will be judged. If yes, then the optimized results are obtained. If no, then the evaluation and reproduction steps are repeated until a certain number of generations, until a defined fitness or until a convergence criterion of the population are reached. In the ideal case, all chromosomes of the last generation have the same genes representing the optimal solution.

Through the process of GA optimization, the stocks are ranked according to the fundamental financial information and price return. Investors can select the top n stocks to construct a portfolio.

3 Experiment Analysis

The daily data used in this study is stock closing price obtained from Shanghai Stock Exchange (SSE) (http://www.sse.com.cn). The sample data span the period from January 2, 2002 to December 31, 2004. Monthly and yearly data in this study are obtained by daily data computation. For simulation, 100 stocks are randomly selected. In this study, we select 100 stocks from Shanghai A share, and their stock codes vary from 600000 to 600100.

First of all, the company financial information as the input variables is fed into the GA to obtain the derived company ranking. This output is compared with the actual stock ranking in terms of APR, as indicated by Equations (5) and (6). In the process of GA optimization, the RMSE between the derived and the actual ranking of each stock is calculated and served as the evaluation function of the GA process. The best chromosome obtained is used to rank the stocks and the top n stocks are chosen for the portfolio. For experiment purpose, the top 10 and 20 stocks are chosen for testing according to the ranking of stock quality using GA. The top 10 and 20 stocks selected by GA can construct a portfolio. For convenience, equally weighted portfolios are built for comparison purpose.

In order to evaluate the usefulness of the GA optimization, we compared the net accumulated return generated by the selected stock from GA with a benchmark. The benchmark return is determined by an equally weighted portfolio of all the stocks available in the experiment. Fig. 2 reveals the results for different portfolios.

From Fig. 2, we can find that the net accumulated return of the equally weighted portfolio formed by the stocks selected by GA is significantly outperformed the benchmark. In addition, the performance of the portfolio of the 10 stocks is better that of the 20 stocks. As we know, portfolio does not only focus on the expected return but also on risk minimization. The larger the number of stocks in the portfolio is, the more flexible for the portfolio to make the best composition to avoid risk. However, selecting good quality stocks is the prerequisite of obtaining a good portfolio. That is, although the portfolio with the large number of stocks can lower the risk to some extent, some bad quality stocks may include into the portfolio, which influences the portfolio performance. Meantime, this result also demonstrates that the portfolio with

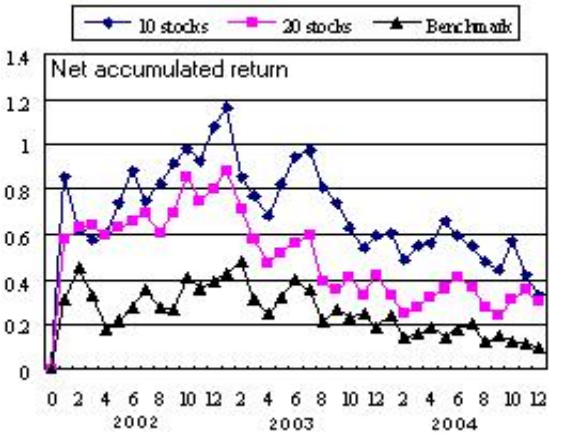

Fig. 2. Accumulated return for different portfolios

the large number of stocks does not necessary outperform the portfolio with the small number of stocks if the investors select good quality stocks. Therefore it is wise for investors to select a limit number of good quality stocks for constructing a portfolio.

4 Conclusions

This study uses genetic optimization algorithm to perform stocks selection for portfolio. Experiment results reveal that the GA optimization approach has shown to be useful to the problem of stock selection, which can help investors select the most valuable stocks for portfolio.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (NSFC No. 70221001); Key Laboratory of Management, Decision and Information Systems of Chinese Academy of Sciences and Strategic Research Grant of City University of Hong Kong (SRG No. 7001677).

References

1. Levin, A.U.: Stock Selection via Nonlinear Multi-factor Models. Advances in Neural Information Processing Systems (1995) 966-972
2. Chu, T.C. Tsao, C.T. Shiue, Y.R.: Application of Fuzzy Multiple Attribute Decision Making on Company Analysis for Stock Selection. Proceedings of Soft Computing in Intelligent Systems and Information Processing (1996) 509-514
3. Zargham, M.R., Sayeh, M.R.: A Web-Based Information System for Stock Selection and Evaluation. Proceedings of the First International Workshop on Advance Issues of E-Commerce and Web-Based Information Systems (1999) 81-83

4. Fan, A., Palaniswami, M.: Stock Selection Using Support Vector Machines. Proceedings of International Joint Conference on Neural Networks 3 (2001) 1793-1798
5. Lin, L., Cao, L., Wang, J., Zhang, C.: The Applications of Genetic Algorithms in Stock Market Data Mining Optimization. In: Zanasi, A., Ebecken, N.F.F., Brebbia, C.A. (Eds.): Data Mining V, WIT Press (2004)
6. Chen, S.H. Genetic Algorithms and Genetic Programming in Computational Finance. Kluwer Academic Publishers, Dordrecht (2002)
7. Thomas, J., Sycara, K.: The Importance of Simplicity and Validation in Genetic Programming for Data Mining in Financial Data. Proceedings of the Joint AAAI-1999 and GECCO-1999 Workshop on Data Mining with Evolutionary Algorithms (1999)
8. Holland, J. H.: Genetic Algorithms. Scientific American 267 (1992) 66-72
9. Goldberg, D.E.: Genetic Algorithm in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, MA (1989)