2C4-IOS-3c-2

A Multi-objective Genetic Model for Stock Selection

Shin-Shou Chen[1] Chien-Feng Huang[2] Tzung-Pei Hong[*1, 2]

[1]National Sun Yat-sen University, Taiwan

[2]National University of Kaohsiung, Taiwan

Stock selection has long been recognized as a challenging and important task in finance. Recent advances in machine learning and data mining are leading to significant opportunities to solve these problems more effectively. In this study, we enrich our previous work for stock selection using single-objective genetic algorithms (SOGA) by extending it to the multi-objective GA (MOGA). In our previous work, we devised a stock scoring mechanism to rank and select stocks to form a portfolio, and we employed the SOGA for optimization of model parameters and feature selection for input variables to the model. In this work, we show how our MOGA models outperform the benchmark and improve upon our previous SOGA-based methods. Based on the promising results, we expect this MOGA methodology to advance the current state of research in soft computing for the real-world stock selection applications.

1. Introduction

In the past decade, several soft computing models have been developed for financial applications, including artificial neural networks (ANNs), support vector machines (SVMs), evolutionary algorithms (EAs) as well as fuzzy inference models. In the particular area of stock selection, Quah and Srinivasan [Quah 1999] studied an ANN stock selection system to choose stocks that are top-ranked performers. They showed their model outperformed the benchmark in terms of compounded actual returns overtime. Chapados and Bengio [Chapados 2001] also trained neural networks for the prediction of asset behavior and decision-making for asset allocation. Typically, these models suffered from overfitting problem and convergence of solutions to local optima.

For portfolio optimization, Kim and Han [Kim 2000] proposed a genetic algorithm (GA) [Holland 1975] approach to feature discretization and the determination of connection weights for ANNs to predict the stock index. They showed that their approach was able to reduce the dimension of variables and the prediction performance was enhanced. In addition, Becker *et al.* [Becker 2006] explored various single-objective fitness functions for GP to construct stock selection models for particular investment specifics with respect to risk. More recently, Huang *et al.* [Huang 2011] proposed a hybrid fuzzy-GA model for stock selection. Based on several statistical tests, Huang *et al.* showed their model can outperform the benchmark significantly. In a nutshell, these GP/GA-based models rank stocks from high to low according to a pre-defined single objective function.

In some financial applications, however, various objectives may impose challenges to the researchers because these objectives are usually competing and the trade-off of selecting distinct solutions one way or another is typically contingent upon one's particular goal. In the general research area of multi-objective optimization (MOO), Hassan and Clack [Hassan 2009] provided some empirical results on the robustness of multiple objective genetic programming (MOGP). They studied two mechanisms — mating restriction and diversity preservation — to determine which leads to more robust solutions. In addition, Sülflow *et al.* [Sülflow 2007] studied multi-objective optimization for high dimensional spaces and presented the pros and cons of existing approaches. More recently, Lohpetch and Corne [Lohpetch 2011] found that multi-objective strategies provide more robustness in outperforming the buy-and-hold strategy for financial trading.

In this work, we extend our previous single objective GA-based model (SOGA) in [Huang 2011] to a multi-objective GA model (MOGA) for the task of stock selection. We will provide the extended formulation for the fitness function to calculate stock scores. Based on the scores calculated, top-ranked stocks are then selected. We will show that our new scheme does improve upon our previous one.

This paper is organized into five sections. Section 2 outlines the methods employed in our study. In Section 3, we present the experimental design and empirical results are reported and discussed. Section 4 concludes this paper with future research directions.

2. Methodology

This section first describes the relevant components for our stock scoring model. Afterwards, model optimization by the GA and the corresponding extension to MOGA will be discussed.

2.1 Stock scoring via fundamental variables

In this study, we are concerned with the relative quality of stocks described by the fundamental variables, including firms' share price rationality, growth, profitability, liquidity, efficiency, and leverage attributes. In general, these fundamental variables can be used to determine the value of a stock, defined by the score assigned by our proposed model. Our objective of this scoring model is to imply stocks of higher scores to possess higher potential in future price advancement. Based on these scores one can then rank various stocks and top-ranked stocks are picked to construct the portfolio.

Contact: Tzung-Pei Hong, Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, R.O.C., tphong@nuk.edu.tw

In this study, we employ a straightforward linear model using these fundamental variables to score stocks [Huang 2011]. Let $X_{i,j,t}$ denote the score of stock i assigned by variable j at time t, where $X_{i,j,t}$ depends on the value of variable j, $v_{i,j,t}$, for stock i at time t. For instance, in the area of value investing, if the variable is the price-to-book ratio (P/B ratio), a smaller P/B ratio tends to imply the stock's higher potential of price increase in the future. On the contrary, if the variable is return-on-assets (ROA), a larger value of ROA usually implies the stock's higher potential of price increase in the future [Huang 2011].

In [Huang 2011], we proposed to use the value of variable j for stock i to determine the individual score assigned to it at time t:

$$X_{i,j,t} = \rho_{i,j,t},$$

where $\rho_{i,j,t} \in N$ is the ranking of stock i with respect to variable j at time t. Here we denote a stock sorting indicator Ij for variable j and consider two cases for the stock sorting scheme:

$$(1)\ I_j = 0\text{: } \rho_{i,j,t} \geq \rho_{k,j,t} \text{ iff } v_{i,j,t} \geq v_{k,j,t} \text{ for } i \neq k.$$
$$(2)\ I_j = 1\text{: } \rho_{i,j,t} \geq \rho_{k,j,t} \text{ iff } v_{i,j,t} \leq v_{k,j,t} \text{ for } i \neq k.$$

In addition, let W_j denote the weight of the j-th variable. Then the total score of stock i at time t, $y_{i,t}(\boldsymbol{W})$, can be defined as:

$$y_{i,t}(W) = \sum_j W_j X_{i,j,t}, \tag{1}$$

where $\boldsymbol{W}$ denotes the vector of the weights of the input features used by the stock scoring model.

Given the scores for all stocks, the ranking of a stock can be defined as:

$$\alpha_{i,t}(W) = \rho(y_{i,t}(W)), \tag{2}$$

where $\rho(\cdot)$ is a ranking function so that $\alpha_{i,\,t} \in N$ is the ranking of stock i at time t, and $\alpha_{i,\,t} \geq \alpha_{j,\,t}$ iff $y_{i,\,t} \geq y_{j,\,t}$.

The task of stock selection can be achieved using these rankings whereby top-ranked m stocks (stocks corresponding to the top m α's) are selected as components of a portfolio. The performance of a portfolio can be evaluated by averaging the actual returns of the stocks in the portfolio, which is defined as:

$$\overline{R_t} = \frac{1}{m}\sum_{i=1}^{m} R_t(s_{i,t}), \tag{3}$$

where $s_{i,t}$ is the i-th ranked stock at time t; $Rt(\cdot)$ is the actual return for a stock at time t and $\overline{R_t}$ is the average return over all the m stocks in the portfolio at time t.

In this study we will use the cumulative total (compounded) return, R_c, to evaluate the performance of a stock selection model, where R_c is defined as the product of the average yearly return, $\overline{R_t}$, of the stocks in a portfolio over n consecutive years as:

$$R_c = \prod_{t=1}^{n} \overline{R_t}. \tag{4}$$

2.2 Model optimization by single-objective genetic algorithms

The performance of the stock selection model is determined by the set of input features $\boldsymbol{F}$, the set of stock sorting indicators $\boldsymbol{I}$, the weights of the fundamental variables $\boldsymbol{W}$. Therefore, the selection of optimal subsets of features $\boldsymbol{F}$, and the optimization of $\boldsymbol{I}$ and $\boldsymbol{W}$ shall be critical to the success of the stock selection model. In [Huang 2011], we used GA for simultaneous optimization with respect to these tasks. Here we provide the description for the relevant GA-based optimization scheme for our stock selection model.

In the overall encoding design, the composition of a chromosome is devised to consist of three portions — the candidate set of features $\boldsymbol{F}$, the stock sorting indicators $\boldsymbol{I}$ and the weighs $\boldsymbol{W}$. In this study, the binary coding scheme is used to represent a chromosome. In Fig. 1, loci b_f^1 through b_f^n represent candidate features 1 through n, respectively. For these features, allele '1' or '0' corresponds to the feature being selected or not. Loci b_i^1 through b_i^n represent the sorting indicators, where 0 represents the variable being used for case (1) of our stock sorting scheme, and 1 represents case (2), respectively. On the right-hand side of Fig. 1 is the encoding of the set of parameters $\boldsymbol{W}$. Fig. 2 shows the detailed binary encoding for the weight of each individual variable where the value of Wi (the weight for variable i) is encoded by loci b_{Wi}^1 through $b_{Wi}^{n_i}$.

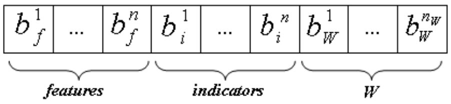

Figure 1. Chromosome encoding

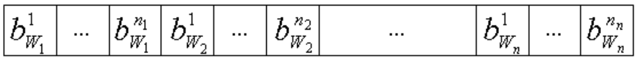

Figure 2. Detailed encoding of the weighting terms

In this coding scheme, the portion in the chromosome representing the genotypes of parameter Wi's is to be transformed into the phenotype by Eq. (5) for further fitness computation. The precision representing each parameter depends on the number of bits used to encode it in the chromosome, which can be determined as follows:

$$y = \min{}_y + \frac{d}{2^l - 1} \times (\max{}_y - \min{}_y), \tag{5}$$

where y is the corresponding phenotype for the particular parameter; $\min_y$ and $\max_y$ are the minimum and maximum of the parameter; d is the corresponding decimal value, and l is the length of the block used to encode the parameter in the chromosome.

With this encoding scheme, in this study we define the fitness of a chromosome as the function of the annualized return and the risk of a portfolio as follows:

$$fitness = \sqrt[n]{R_c} \,/\, \sigma, \tag{6}$$

where Rc is the cumulative total return computed by Eq. (4), and σ is the standard deviation of all the average yearly returns computed by Eq. (3).

2.3 Model optimization by multi-objective genetic algorithms

In this study we propose to use the Non-dominated Sorting Genetic Algorithms II (NSGA-II) [Deb 2002] for multi-objective optimization with respect to the two seemingly conflicting objectives of return and risk for investment. The goal of NSGA-II is to quickly search for solutions on the Pareto-optimal fronts and maintain diversity among them.

In the NSGA-II setup here, the goal of our stock selection models is to generate portfolios of higher return and lower risk. One can employ these two objectives to determine the dominance among solutions in a population. The solutions are then ranked according to the Pareto-front on which they reside. That is, the solutions on the first non-dominating front are assigned a rank of one; the solutions on the second non-dominating front are assigned rank of two, and so on. Top-ranked solutions are then selected for reproduction for the next generation.

In order to maintain the diversity among solutions on the same Pareto front, crowding distance for each solution is computed. The crowding distance measures the distance of the biggest cuboid containing the two neighboring solutions on the same non-dominating front. On the same non-dominating front, solutions with larger crowding distance are more likely to be selected for reproduction for the next generation.

For the overall selection procedure, solutions with higher ranks get selected for reproduction; for solutions of the same rank, the ones with larger crowding distance are then selected. The children and parent population are then combined together for elitism and the mechanism of non-dominating sorting is applied on the new population repeatedly.

3. Empirical Results

We use the constituent stock of the 200 largest market capitalizations listed in the Taiwan Stock Exchange as the investment universe. The yearly financial statement data and stock returns used for this research are retrieved from the TEJ (Taiwan Economic Journal Co. Ltd., http://www.tej.com.tw/) database for the period of time from 1987 to 2009. For the choice of fundamental variables, early studies indicated that several financial ratios play key roles in future stock returns. Table 1 provides fourteen attributes that are to be employed for this study. For each year, investable stocks are described by these fourteen financial ratios and their historical returns are provided.

In order to provide a statistical validation for our proposed models, we split the data into two parts: the first n years of the data is used to train the models and the remaining data is used to for validation. For instance, the data of 1987 through 1990 can be used for training, and the data for the remaining years (1991-2009) can be used to test the models learned from the training phase.

In Table II, we display the results for three cases only in the testing phase because one is usually concerned with how the models perform in testing, including the annualized returns and Sharpe ratios for the benchmark (all the 200 stocks are used as a portfolio to compute the statistics), and for the top 10% of the 200 stocks selected by our single-objective GA (SOGA) and MOGA models. An inspection on the means of annualized returns shows that the SOGA model outperforms the benchmark in 11 out of 16 cases, and the MOGA model outperforms the benchmark in 15 out of 16 cases. Furthermore, an inspection on the means of the Sharpe ratios shows that the SOGA model outperforms the benchmark in 12 out of 16 cases, and the MOGA model outperforms the benchmark in 15 out of 16 cases. Therefore, using the annualized returns and Sharpe ratios, one can see that the MOGA model further improves the SOGA model, and both of them outperform the benchmark, as well.

To further illustrate the performance discrepancy of the models and the benchmark, Table III also displays the results for selecting top 20% of the 200 stocks by the SOGA and MOGA models. The results show that the means of annualized returns by the SOGA model outperforms the benchmark in 13 out of 16 cases, and the MOGA model outperforms the benchmark in 15 out of 16 cases. In terms of the Sharpe ratios, the results show that the SOGA model outperforms the benchmark in 13 out of 16 cases, and the MOGA model outperforms the benchmark in 14 out of 16 cases. Therefore, one can again see that the MOGA model further improves the SOGA model, and both of them outperform the benchmark.

4. Conclusions

In this paper we present a multi-objective GA methodology for stock selection to improve upon our previous single-objective GA models. Based on the extended MOGA-based stock scoring mechanism, top-ranked stocks can be selected more effectively as components in a portfolio. We have evaluated the proposed models statistically and showed that our proposed MOGA model does improve our previous one and outperform the benchmark significantly. Therefore, we expect this MOGA methodology to advance the research in computational finance and provide a promising solution for stock selection.

References

[Bauer 2004] R. Bauer, N. Guenster, R. Otten, “Empirical evidence on corporate governance in Europe: the effect on stock returns, firm value and performance,” Journal of Asset Management, vol. 5(2), 91-104, 2004.

[Becker 2006] Y. Becker, P. Fei, A. Lester, “Stock selection – An innovative application of genetic programming methodology”, in: R. Riolo, T. Soule, B. Worzel (Eds.), Genetic Programming Theory and Practice IV, Genetic and Evolutionary Computation, vol. 5, Springer, Ann Arbor, Michigan, pp. 315– 334, Chapter 12, 2006.

[Chapados 2001] N. Chapados, Y. Bengio, “Cost functions and model combination for VaR-based asset allocation using neural networks”, IEEE Transactions on Neural Networks, vol. 12, pp. 890–906, 2001.

[Carnes 2006] T. A. Carnes, “Unexpected Changes in quarterly financial-statement line items and their relationship to stock prices,” Academy of Accounting and Financial Studies Journal, vol. 10(3), 99-116, 2006.

Table I. Variables used in the stock selection model

Ratios	Description	Ref.
PE ratio	Price-to-earnings ratio = share price / earnings per share	[Mukherji 1997]
PB ratio	Price-to-book ratio = share price / book value per share	[Mukherji 1997]
PS Ratio	Price-to-sales ratio = share price / sales per share	[Mukherji 1997]
ROE	Return on equity (after tax) = net income after tax / shareholders’ equity	[Omran 2004][Bauer 2004]
ROA	Return on asset (after tax) = net income after tax / total assets	[Omran 2004]
OPM	Operating profit margin = operating income / net sales	[Soliman 2008]
NPM	Net profit margin = net income after tax / net sales	[Bauer 2004]
DE ratio	Debt-to-equity ratio = total liabilities / shareholders’ equity	[Omran 2004]
CR	Current ratio = current assets / current liabilities	[Omran 2004]
QR	Quick ratio = quick assets / current liabilities	[Omran 2004]
ITR	Inventory turnover rate = cost of goods sold / average inventory	[Omran 2004]
RTR	Receivables turnover rate = net credit sales / average accounts receivable	[Carnes 2006]
OIG	Operating income growth rate = (operating income at the current year – operating income at the previous year) / operating income at the previous year	[Ikenberry 1993]
NIG	Net income growth rate = (net income after tax at the current year – net income after tax at the previous year) / net income after tax at the previous year	[Sadka 2009]

Table II. Statistics of the benchmark, SOGA and MOGA models for 20 stocks

Testing period	Annualized benchmark return	Mean of annualized SOGA model returns	Mean of annualized MOGA model returns	Testing period	Annualized benchmark Sharpe Ratio	Mean of Sharpe ratios for SOGA model	Mean of Sharpe ratios for MOGA model
1992-2009	2.8425	5.5198	4.2599	1992-2009	0.2345	0.2757	0.2481
1993-2009	3.9115	8.3673	6.8789	1993-2009	0.2690	0.3449	0.3094
1994-2009	0.7016	1.9416	4.1918	1994-2009	0.1655	0.1982	0.2226
1995-2009	3.8463	6.8494	7.1536	1995-2009	0.2676	0.3189	0.2935
1996-2009	2.3489	6.2107	8.6010	1996-2009	0.2154	0.3072	0.3208
1997-2009	-1.9077	-1.5815	1.9078	1997-2009	0.0366	0.0693	0.1931
1998-2009	0.2147	-0.1911	0.6143	1998-2009	0.1313	0.1253	0.1536
1999-2009	0.0694	0.2305	3.5907	1999-2009	0.1301	0.1486	0.2641
2000-2009	1.4078	-1.3759	4.9229	2000-2009	0.1921	0.0856	0.3279
2001-2009	8.5513	8.0507	11.8293	2001-2009	0.5635	0.5712	0.6382
2002-2009	7.4737	8.2344	10.0926	2002-2009	0.4873	0.5721	0.5568
2003-2009	6.3206	5.5574	7.8145	2003-2009	0.4116	0.3967	0.4396
2004-2009	7.0959	8.0985	8.6456	2004-2009	0.4349	0.5269	0.4934
2005-2009	7.1129	8.3467	7.7138	2005-2009	0.4169	0.5341	0.4242
2006-2009	4.8263	3.6790	3.8344	2006-2009	0.3088	0.2990	0.2456
2007-2009	-5.9271	-4.0221	-5.2485	2007-2009	-0.2695	-0.1585	-0.2361

Table III. Statistics of the benchmark, SOGA and MOGA models for 40 stocks

Testing period	Annualized benchmark return	Mean of annualized SOGA model returns	Mean of annualized MOGA model returns	Testing period	Annualized benchmark Sharpe Ratio	Mean of Sharpe ratios for SOGA model	Mean of Sharpe ratios for MOGA model
1992-2009	2.8425	8.0713	5.5844	1992-2009	0.2345	0.3481	0.2928
1993-2009	3.9115	9.4235	9.7197	1993-2009	0.2690	0.3784	0.3834
1994-2009	0.7016	6.4717	4.8268	1994-2009	0.1655	0.3073	0.2658
1995-2009	3.8463	9.8716	9.8387	1995-2009	0.2676	0.3871	0.3830
1996-2009	2.3489	9.5264	6.7784	1996-2009	0.2154	0.3744	0.3114
1997-2009	-1.9077	-1.2405	1.1269	1997-2009	0.0366	0.0789	0.1695
1998-2009	0.2147	0.3348	1.7387	1998-2009	0.1313	0.1488	0.1964
1999-2009	0.0694	0.6420	2.7112	1999-2009	0.1301	0.1630	0.2378
2000-2009	1.4078	-0.4866	5.5264	2000-2009	0.1921	0.1255	0.3548
2001-2009	8.5513	8.7302	11.1349	2001-2009	0.5635	0.5729	0.6500
2002-2009	7.4737	8.7743	9.9675	2002-2009	0.4873	0.5699	0.5732
2003-2009	6.3206	7.1764	7.9783	2003-2009	0.4116	0.4699	0.4752
2004-2009	7.0959	8.6015	8.9944	2004-2009	0.4349	0.5239	0.4919
2005-2009	7.1129	7.5627	7.5431	2005-2009	0.4169	0.4493	0.4158
2006-2009	4.8263	3.3733	6.6457	2006-2009	0.3088	0.2591	0.3533
2007-2009	-5.9271	-6.1225	-6.7939	2007-2009	-0.2695	-0.3469	-0.3468

[Deb 2002] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," Transactions on Evolutionary Computation, vol. 6(2), 182-197, 2002

[Holland 1975] J. H. Holland, "Adaptation in natural and artificial systems," University of Michigan Press, Ann Arbor, Michigan, 1975.

[Hassan 2009] G. Hassan, C. Clack, "Robustness of Multiple Objective GP Stock-Picking in Unstable Financial Markets", GECCO'09 Proceedings of the 11th Annual conference on Genetic and evolutionary computation, pages 1513-1520, 2009.

[Huang 2011] Huang, C.-F., Chang, C.-H., Chang, B. R. and Cheng, D.-W. (2011). "A Study of a Hybrid Evolutionary Fuzzy Model for Stock Selection," Proceeding of the 2011 IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, June 27-30, pp. 210-217, 2011.

[Ikenberry 1993] D. Ikenberry, J. Lakonishok, "Corporate governance through the proxy contest: evidence and implications," Journal of Business, vol. 66(3), 405-435, 1993.

[Kim 2000] K. J. Kim, I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", Expert Systems with Applications, vol.19, pp. 125–132, 2000.

[Lohpetch 2011] D. Lohpetch, D. Corne, "Multiobjective Algorithms for Financial Trading", Evolutionary Computation (CEC), 2011.

[Mukherji 1997] S. Mukherji, M. S. Dhatt, Y. H. Kim, "A fundamental analysis of Korean stock returns," Financial Analysts Journal, vol.53(3), pp. 75-80, 1997.

[Omran 2004] M. Omran, "Linear versus non-linear relationships between financial ratios and stock returns: empirical evidence from Egyptian firms," Review of Accounting and Finance, vol. 3(2), 84-102, 2004.

[Quah 1999] T. S. Quah, B. Srinivasan, "Improving returns on stock investment through neural network selection", Expert Systems with Applications, vol.17, pp. 295–301, 1999.

[Sülflow 2007] A. Sülflow, N. Drechsler and R. Drechsler, "Robust Multi-Objective Optimization in High Dimensional Spaces", Lecture Notes in Computer ScienceVolume 4403, pp 715-726, 2007.

[Soliman 2008] M. T. Soliman, "The use of DuPont analysis by market Participants," The Accounting Review, vol. 83(3), 823-853, 2008.

[Sadka 2009] G. Sadka, R. Sadka, "Predictability and the earnings-returns relation," Journal of Financial Economics, vol. 94(1), 87-93, 2009.