Evaluating Efficiency of Index Fund Selections Over the Fund's Future Period

Yukiko Orito[1] Manabu Takeda[2] Kiyoaki Iimura[2] Genji Yamazaki[2]

[1]Ashikaga Institute of Technology, 268-1, Ohmae-cho, Ashikaga, Tochigi 326-8558, Japan
[2]Tokyo Metropolitan Institute of Technology, 6-6, Asahigaoka, Hino, Tokyo 191-0065, Japan

Abstract

It is well known that index fund selections are important for hedge trading investment in a stock market. "*Selection*" means selection of a certain number of listed companies from a specific market. The index funds are constructed by application of a genetic algorithm. The main purpose of this paper is to evaluate the efficiency of the index fund selections method over the fund's future period. This method is examined with numerical experiments applied to the Tokyo Stock Exchange. The results show that with this method, index funds work well for forecasting over a future period when a market index has been on a downward trend.

Keywords: Index Fund Selections, Market Index, Genetic Algorithms.

1. Introduction

"*Index funds*" have been used very extensively for hedge trading, which is the practice of offsetting the price risk on any stock market position by taking an equal, but opposite position in a futures market. At the stock selection level, suppose that we select n listed companies on a stock market and invest in a certain number of stocks of each of the n companies. It is well known that a group consisting of n companies is very useful for hedge trading if the total return rate of the group follows a similar path to the changing rate of the market index. Then we have the important problem of finding a group consisting of n companies on the market. The coefficient of determination between the total return rate of the group and the changing rate of the market index, R^2 (defined in Section 2), is usually used as a measure of how the return rate follows the changing rate (see, e.g., Downie and Heath [1]). Index fund selections can be viewed as a combinatorial optimization problem. That is to say that it is important to select the n companies on the stock market where the R^2 is high. On the other hand, the index fund requires rebalancing in order to reflect the changes in the composition of the market index over the fund's future period. However, the total price of the index fund is unknown, so the implied cost of rebalancing is uncertain. To construct the group consisting of n companies, we have to make a great investment in rebalancing when the number of companies is large. From a practical viewpoint it is preferred that the group consisting of n companies is constructed such that n is small but the R^2 is high. In this context, Takabayashi [5] reported that for the Tokyo Stock Exchange, consisting of about 1400 listed companies, his method constructed the index funds consisting of $n \approx 440$ companies whose R^2 was about 0.96. We proposed a method consisting of the following two steps:

Step 1. Each listed company on the market is weighted (or assigned a value) according to a "*heuristic rule*" based on a company's volume. A set of N companies is chosen by using this rule.
Step 2. The group of n companies is selected by applying a genetic algorithm (GA) to the N companies.

We applied this GA method to the Tokyo Stock Exchange and reported that it was possible to construct the index funds consisting of $n \approx 100$ companies whose R^2 was about 0.98 [4]. The main purpose of this paper is to evaluate the efficiency of the index fund selections over the fund's future period. The results show that with the GA method, the index funds work well for forecasting over the future period when the market index has been on a downward trend.

2. Preliminaries

Suppose that we invest in a group consisting of n listed companies on a stock market, which starts at $t = 1$ and ends at $t = T$. The t is set on a data basis. It is assumed throughout the paper, the amount of money invested in each company, belonging to the group, is fixed and the group consisting of n companies is selected at $t = T$ before rebalancing. This means that the portfolio is unique for the group. In the field of regression analysis, the coefficient of determination

has often used as a measure of how well an estimated regression fits. As the coefficient approaches 1, the estimated regression fits better (for this, see, e.g., Downie and Heath [1]). By analogy, the coefficient of determination between the return rate of the group and the changing rate of the market index is defined as

$$R^2 = \frac{\left(T\sum_{t=1}^{T} x(t)y(t) - \sum_{t=1}^{T} x(t)\sum_{t=1}^{T} y(t)\right)^2}{\left(T\sum_{t=1}^{T} x^2(t) - \left(\sum_{t=1}^{T} x(t)\right)^2\right)\left(T\sum_{t=1}^{T} y^2(t) - \left(\sum_{t=1}^{T} y(t)\right)^2\right)},$$

where $x(t)$ is the return rate of the index fund between $t=1$ and $t=T$ and $y(t)$ is the changing rate of the market index between $t=1$ and $t=T$. In this paper, the R^2 is adopted as the fitness measure.

3. GA Method

As mentioned in Section 1, the GA method consists of the following two steps.

3.1. Step 1

Suppose that a stock market consists of K listed companies, numbered company 1, company 2, $\cdots$, company K. The first step of the GA method is to choose N companies of K companies according to the heuristic rule. The rule of Orito et al. [4]'s is based on "*volume* (trading volume divided by shares outstanding)." For a company i on the market, the company's value average between $t=1$ and $t=T$ is defined by

$$V_i = \frac{1}{T}\sum_{t=1}^{T} w_i(t)z_i(t),$$

where $w_i(t)$ is the volume of company i at time t and $z_i(t)$ is the stock price of company i at time t. Suppose that the V_i s are assigned to the K companies on the market. Without loss of generality, we can renumber the K companies so that

$$V_1 \geq V_2 \geq \cdots \geq V_i \geq \cdots \geq V_K.$$

We note that the renumbered company i has the i-th high V_i of all companies. Therefore the heuristic rule means that the set of company 1, company 2, $\cdots$, company N is chosen.

3.2. Step 2

The second (final) step is formulated as the problem of finding a group consisting of n companies, such that the R^2 is the highest in R^2 s of all subsets of N companies. It is well known that GAs are useful for such optimization problems (for this, see, e.g., Melanie [2]). Hence, we use a GA for finding the n companies from the N ones. Suppose that the set of N companies was given. The approach of implementation of the GA follows that of Orito et al. [4]. A gene is defined by

$$g_i = \begin{cases} 0 & \text{Company } i \text{ is not an element of } n \text{ companies} \\ 1 & \text{Company } i \text{ is an element of } n \text{ companies} \end{cases} \quad (i = 1,2,\cdots,N)$$

and a chromosome is denoted by $\bar{g} = \{g_1, g_2, \cdots, g_N\}$. The "*fitness value of the GA*" is R^2. The GA is designed as follows:

1. **Beginning.** On the 1st generation of the GA, we randomly generate 100 chromosomes in the initial population.
2. **Evaluation of the fitness value.** Next we select one chromosome which has the highest R^2 in the current population, say $\bar{g}_{\max}$.
3. **Crossover.** After step two which is the evaluation of the fitness value, we generate a random number r in $[0,1]$ for each chromosome in the current population. If the r is less than a crossover rate $Pc = 0.9$, the crossover makes new chromosomes by exchanging the partial structure between the chromosome and another one selected at random. The exchange of the partial structure position is determined at random in both chromosomes.
4. **Mutation.** Following the crossover selection process, we generate a random number r in $[0,1]$ for each chromosome in the current population. If the r is less than a mutation rate $Pm = 0.05$, the mutation makes new chromosomes by replacing the partial structure of the chromosome. The partial structure position is determined at random and replaced 0 with 1 or 1 with 0.
5. **Selection.** After the crossover selection and mutation process, suppose that there are M chromosomes, numbered $\bar{g}_1$ through $\bar{g}_M$, after applying the crossover and the mutation. Let f_i be the R^2 for $\bar{g}_i$ and let p_i be the rate $f_i \Big/ \sum_{j=1}^{M} f_j$. The (cumulative) probability $q_i = \sum_{j=1}^{i} p_j$ is assigned to $\bar{g}_i$. We generate a random number r in $[0,1]$. The $\bar{g}_i$ is selected as an element of the new population when $q_{i-1} < r \leq q_i$. Repeating this procedure, we select 100 chromosomes in the new population. If the highest R^2 in the new

population is less than the R^2 of the $\overline{g}_{\max}$, one of the 100 chromosomes is randomly replaced with the $\overline{g}_{\max}$. On the next generation of the GA, the new population consists of 100 chromosomes that were chosen by this selection process.

6. **Stop.** Finally the GA is broken off on the 100 th generation. If the number of generations is less than 100, the GA goes back to step two which is the evaluation of the fitness value.

On the 100 th generation, the chromosome $\overline{g}_{\max}$ with the highest R^2 is obtained. The companies having $g_i = 1$ of the $\overline{g}_{\max}$, makes up the group consisting of *n* companies. For fixed *N* companies, the GAs are executed 20 times and 20 groups are given. We select one group which has the nearest R^2 to the average of 20 group's R^2 s. This is our index fund by applying the GA method.

4. Numerical Experiments

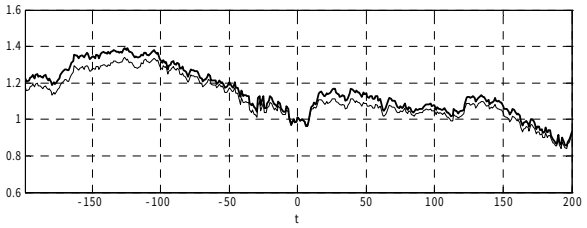
We have applied the GA method to the data periods on the Tokyo Stock Exchange consisting of more than 1500 listed companies. In this section, the results for 27 cases, say Case 1, Case 2, ⋯, Case 27, are shown. For each case, the data period in the numerical experiment is the length on the 400 days, $t \in [-199, 200]$. The period is shifted every 30 days between Mar. 11, 1997 and Dec. 21, 2001.

For each case, we obtain the index fund by applying the GA method to the data from $t = -199$ to $t = 0$ and evaluate the efficiency of the index fund over the fund's future period from $t = 1$ and $t = 200$. In this paper, the former data period $[-199, 0]$ is called a "*past period*" and the latter $[1, 200]$ is called a "*future period*", respectively.

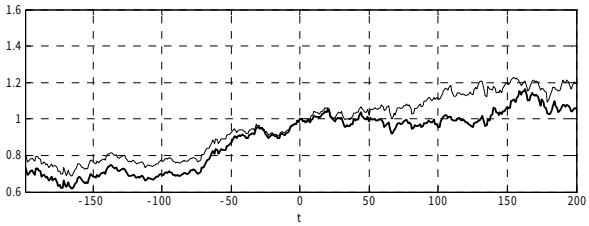
4.1. Main Results

In the past period of each case, the GA method selected the index fund consisting of *n* companies from $N = 200$. For example, the result for Case 1 or Case 13 are shown in Fig. 1 (a) or (b), respectively. The time is the horizontal axis. Let $Q(t)$ be the market index at time t. As the market index of Case 1 or Case 13, the Tokyo Stock Price Index (TOPIX) was used and the movement of TOPIX normalized with $Q(0)$, $Q^*(t) = Q(t)/Q(0)$, is shown by a thin line in Fig. 1 (a) or (b), respectively. On the other hand, let $A(t)$ be the total price of the index fund at time t. The movement of the total price normalized with $A(0)$, $A^*(t) = A(t)/A(0)$, is shown by a thick line in Fig. 1 (a) or (b), respectively.

Fig. 1 (a) suggests that the total price of the index fund follows a similar path to the market index over the past and the future period. Fig. 1 (b) suggests that the total price of the index fund does not follow the market index over the future period. In other words, the index fund obtained by the GA method works well over the fund's future period in Case 1 but does not work over the fund's future in Case 13. What is the main difference between these results? We discuss this problem in the next section.



(a) Case 1



(b) Case 13

Fig. 1: The market index normalized with the value at $t = 0$ (*thin line*) or the total return of the index fund normalized with the value at $t = 0$ (*thick line*) as a function of time t.

4.2. Efficiency of Index Fund Over the Fund's Future

When the market index over a past period is viewed as a time series data, the data can be characterized as a downward trend data or an upward trend data. The difference in the market index for 100 days over the past period, $Q_{error} = Q^*(0) - Q^*(-100)$, in this paper it is called the "*market index distance error*". From Fig. 1 (a), $Q_{error} = -0.2675$. When the market index distance error is $Q_{error} \le 0$, the market index has been on a downward trend over the past period. On the other hand, in the Case 13, the market index over the past period has been on an upward trend. From Fig. 1 (b), $Q_{error} = +0.2247$. For all 27 cases, we have had the coefficients of determination of the index

funds, R^2 . The market index distance error is the horizontal axis in Fig. 2. The R^2 over the past period of each index fund is plotted by "o" in Fig. 2. The R^2 over the future period is plotted by "$*$" in Fig. 2. This means that one index fund has two kinds of coefficients of determination as a function of the market index distance error. One is for a past period and the other is for a future period.

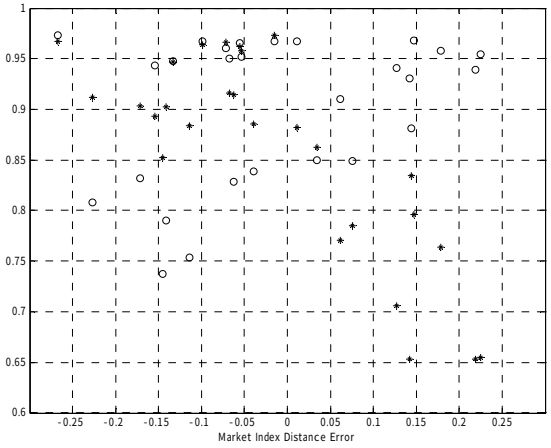


Fig. 2: The coefficients of determination over the past period ("o") and the future period ("$*$") of the index fund as a function of the market index distance error. The negative error means that the market index has been on a downward trend and the positive error means that the market index has been on an upward trend.

We focus on the case studies of the negative market index distance error. When the market index has been on a downward trend, there are 9 case studies whose coefficient of determination over the past period (plotted by "o") is more than 0.9. In these case studies, the coefficient of determination over the future period (plotted by "$*$") is more than 0.9 except in one case. This means that the index funds work well over the fund's future period. On the other hand, there are 7 case studies whose coefficient of determination over the past period is less than 0.9. In these 16 case studies of the negative market index distance error, we adopt the new measure. The error of values between the return rate of the index fund and the market index, $A^*(-100)-Q^*(-100)$, is the horizontal axis in Fig. 3. The R^2 which is more than 0.9 is plotted by "o" in Fig. 3. The R^2 which is less than 0.9 is plotted by "+" in Fig. 3.

The figure suggests that the coefficient of determination of the index fund depends on the error of values between the index fund and the market index. From the results the index funds obtained by the GA method work well over the fund's future period when the market index has been on a downward trend and the coefficient of determination over the past period is more than about 0.9.

When the market index has been on an upward trend, however, the index funds do not work over the fund's future period. This merits future study.

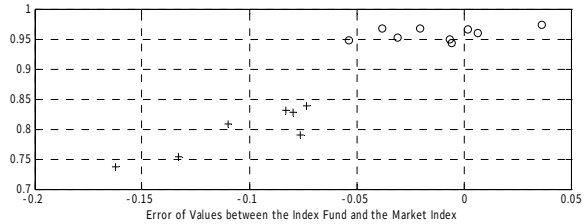


Fig. 3: The coefficient of determination which is more than 0.9 ("o") or which is less than 0.9 ("+") of the index fund as a function of the error of values between the index fund and the market index.

5. Concluding Remarks

We selected index funds on the Tokyo Stock Exchange and evaluated the efficiency of the index funds over the fund's future period. The numerical experiments showed that the index funds worked well over the fund's future when the market index had been on a downward trend. To analyze the problems experienced in the upward trend remains to be seen. This will be revealed in future studies.

6. References

[1] N.M. Downie and R.W. Health, "Basic Statistical Methods," *Harper and Row*, 1983.

[2] M. Melanie, "An Introduction to Genetic Algorithms," *Tokyo Denki University Press*, 1997.

[3] Y. Orito, T. Motoyama, and G. Yamazaki, "Index Fund Selections with GAs and Classifications Based on Turnover," *IEEJ Transactions of the Institute of Electrical Engineers of Japan*, Vol. 124-10, pp. 2014-2018, 2004.

[4] Y. Orito, H. Yamamoto, and G. Yamazaki, "Index Fund Selections with Genetic Algorithms and Heuristic Classifications," *Journal of Computers and Industrial Engineering*, Vol. 45, pp. 97-109, 2003.

[5] A. Takabayashi, "Selections and Rebalancing of Funds with Genetic Algorithms" (in Japanese), *Proc. of the 1995 Winter Conference on Japanese Association of Financial Econometrics and Engineering*, 1995.

[6] H. Tsuda, "Statistics for Assets" (in Japanese), *Asakura Syoten*, 1997.