Feature Selection and Parameter Optimization of a Fuzzy-based Stock Selection Model Using Genetic Algorithms

Article *in* International Journal of Fuzzy Systems · March 2012

CITATIONS
34

READS
416

4 authors, including:

Chien-Feng Huang
National University of Kaohsiung
72 PUBLICATIONS **603** CITATIONS
SEE PROFILE

Chih-Hsiang Chang
National University of Kaohsiung
39 PUBLICATIONS **439** CITATIONS
SEE PROFILE

Feature Selection and Parameter Optimization of a Fuzzy-based Stock Selection Model Using Genetic Algorithms

Chien-Feng Huang, Bao Rong Chang, Dun-Wei Cheng, and Chih-Hsiang Chang

Abstract

In the areas of investment research and applications, feasible quantitative models include methodologies stemming from soft computing for prediction of financial time series, multi-objective optimization of investment return and risk reduction, as well as selection of investment instruments for portfolio management, etc. Among all these, stock selection has long been identified as a challenging and important task. This line of research is highly contingent upon reliable stock ranking for successful portfolio construction. Recent advances in machine learning and data mining are leading to significant opportunities to solve these problems more effectively. In this study, we aim at developing a methodology for effective stock selection using fuzzy models as well as genetic algorithms (GA). We first devise a stock scoring mechanism using fundamental variables and apply fuzzy membership functions to re-scale the scores properly. The scores are then used to obtain the relative rankings of stocks and top-ranked stocks can be selected to form a portfolio. On top of the stock scoring model, we employ GA for optimization of model parameters and feature selection for input variables simultaneously. We will show that the investment returns provided by our proposed methodology significantly outperform the benchmark. Based upon the promising results obtained, we expect that this hybrid fuzzy-GA methodology can advance the research in soft computing for finance and provide an effective solution to stock selection in practice.

Keywords: Feature selection, fuzzy models, genetic algorithms, model validation, parameter optimization, stock selection

1. Introduction

Stock selection has long been identified as a challenging and important research area in finance. The success of this task is highly contingent upon reliable selection of stocks that provide above-average return on investment in the future. Recent advances in computational intelligence and data mining are leading to significant opportunities to solve these problems more effectively. Feasible quantitative models include methodologies stemming from soft computing [1] for prediction of financial time series, multi-objective optimization of expected investment return and risk reduction, and portfolio management – selection of investment instruments based on asset ranking using a variety of input variables and historical data, etc. [2-3]. All these research efforts were in an attempt to facilitate the task of decision-making for investment.

In the research area of stock selection and portfolio optimization, several machine learning methodologies have been developed, including artificial neural networks (ANNs), support vector machines (SVMs), evolutionary algorithms (EAs) as well as fuzzy inference models. Quah and Srinivasan [4] studied an ANN stock selection system to choose stocks that are top-ranked performers. They showed their proposed model outperformed the benchmark model in terms of compounded actual returns overtime. Chapados and Bengio [5] also trained neural networks for estimation and prediction of asset behavior in order to facilitate decision-making in asset allocation. Although these models worked in some applications, they often suffer from the overfitting problem and may tend to fall into a local optimum.

For portfolio optimization, Kim and Han [6] proposed a genetic algorithm (GA) approach to feature discretization and the determination of connection weights for ANNs to predict the stock price index. They suggested that their approach was able to reduce the numbers of attributes and the prediction performance was enhanced. In addition, Caplan and Becker [7] employed genetic programming (GP) to develop a stock ranking model for the high technology manufacturing industry in the U.S. More recently, Becker, Fei and Lester [8] explored various single-objective fitness functions for GP to construct stock selection models for particular investment specifics with respect to risk. In a nutshell, these GP-based models rank stocks from high to low according to a pre-defined objective function.

Corresponding Author: Bao Rong Chang is with the Department of Computer Science and Information Engineering, National University of Kaohsiung, No. 700, Kaohsiung University Rd., Nanzih District, Kaohsiung 811, Taiwan.
E-mail: brchang@nuk.edu.tw
Manuscript received 10 June 2011; revised 8 Oct. 2011; accepted 6 Dec. 2011.

In the area of fuzzy applications in finance, earlier work includes, for instance, Chu *et al*.'s fuzzy multiple attribute decision analysis to select stocks for portfolio construction [9]. Analogously, Zargham and Sayeh [10] employed a fuzzy rule-based system to evaluate a set of stocks for the same purpose. Although these fuzzy approaches denote early efforts in employing computational intelligence for financial applications, they usually lack sufficient learning ability. However, more recently, Chang [11-12] studied a hybrid neuro-fuzzy inference system for the forecast of financial time series. Chang showed that this hybrid model is able to improve the predictive accuracy of irregular non-periodic short-term time series forecast.

Despite the promising performance of the aforementioned approaches in finance, their success is highly contingent upon the input variables (features) to the model. Yang and Honavar [13] indicated that several classification issues are determined by the choice of features that describe given patterns presented to a classifier, such as the classification accuracy of the learned classifier, the computational overhead required for learning a classification function, the number of training examples needed for learning, and the cost associated with the features. Therefore, the goal of feature selection aims to identify useful, non-redundant subsets of features for a given data mining or machine learning task.

In addition, since the variables relevant to the machine learning models usually consist of not only the features but also the models parameters, it is expected that a successful model along this line of research shall take into consideration these two issues simultaneously. Therefore, in this study, we devise a hybrid fuzzy-GA stock selection model for this task, where the GA is used for feature selection and parameter optimization of fuzzy models, simultaneously. Furthermore, in our proposed scheme, feature selection depends on the output of the stock selection model. As a result, our hybrid model can be categorized as a wrapper approach [14-15], as opposed to the filter approach. The wrapper approach for feature selection is employed in this study because of its improved performance over the filter approach [14-18]. In essence, the optimization method we adopted here is very similar to that proposed by Huang and Wang [16], yet we will demonstrate our main contribution lies in a proper setup that successfully applied this hybrid methodology to stock selection and verified its effectiveness with several statistical tests for practical applications, which has not been treated extensively in the machine learning field.

In a nutshell, we will conduct the stock scoring mechanism using fuzzy membership functions with the GA-optimized model parameters and subsets of features. Based on the scores calculated, top-ranked stocks are then chosen for portfolio construction. We will report the portfolios constructed by our proposed scheme substantially outperform the benchmark over the long period of time.

This paper is organized into five sections. Section 2 outlines the methods employed in our study. In Section 3 we describe the research data used in this study. In Section 4, we present the experimental design and empirical results are reported and discussed. Section 5 concludes this paper with future research directions.

2. Methodology

This section first describes our proposed scheme to score stocks using fundamental variables and fuzzy membership functions. Afterwards, model optimization, i.e., parameter optimization and feature selection, by the GA will be discussed.

A. *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 developing this scoring model is to imply that stocks of higher scores shall bear 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.

In this study, we first propose a straightforward linear model using these fundamental variables to score stocks. 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 value (P/B ratio), a smaller P/B value tends to imply the stock's higher potential of price increase in the future [19]. 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.

Therefore, we propose to sort the stocks according to their values of variable j and the individual score assigned to stock i at time t is:

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

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 I_j for variable j and consider two cases for the stock sorting scheme:

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

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

In addition, let W_j denote the weight of the j-th vari-

able. Then the total score of stock i at time t, $S_{i,t}$, can be defined as

$$S_{i,t} = \sum_j W_j X_{i,j,t}. \quad (1)$$

In this study, to allow more flexibility for our model, we introduce a fuzzy membership scheme as an extension of the scoring mechanism, which will be discussed next.

B. *Fuzzy model for stock scoring*

To improve the accuracy of the stock scoring model, the fuzzy membership function is applied. This treatment is to increase the flexibility of the model so that the stock scores calculated by the GA may be adjusted properly to reflect the actual potential of stock price increase in the future. This adjusted score, denoted by $y_{i,t}$, is determined by the score calculated by (1), $S_{i,t}$, and the fuzzy membership function, $\mu(x)$.

More specifically, in the fuzzy stock scoring model, the score for stock i at time t, $y_{i,t}(W, \theta) \in R$, is defined as

$$y_{i,t}(W, \theta) = S_{i,t}(W, \theta) * \mu(S_{i,t}(W, \theta)),$$

where W denotes the vector of the weights of the input features (i.e., the fundamental variables) used by the stock scoring model and θ denotes the set of the parameters used by the fuzzy membership function.

In this study, we compare three commonly used fuzzy membership function — the triangle, trapezoid and Gaussian functions. As an illustration, Figure 1 displays the triangular membership function as shown below, where the shape of the fuzzy model is determined by parameters a, b, and c. In this study, these parameters will be optimized by the GA based on the given fitness function.

$$\mu_A(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a < x \le b \\ \dfrac{c-x}{c-b}, & b < x \le c \\ 0, & c < x \end{cases}$$

Our experimental results so far show that the triangle and trapezoid functions exhibited similar, better performance than the Gaussian function. The reason is that the tip and the width of the triangle and trapezoid functions are adjustable, whereas in the Gaussian function, only the width can be adjusted. Therefore, in this paper, we will only demonstrate the experimental results using the triangular membership function.

It is worthwhile to mention that the scores calculated for the stocks may not necessarily represent the precise values of various stocks. Rather, they can serve as surrogates for the actual quality to imply the relative rankings of the stocks. Thus given the scores for all stocks, the ranking of a stock can be defined as:

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

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

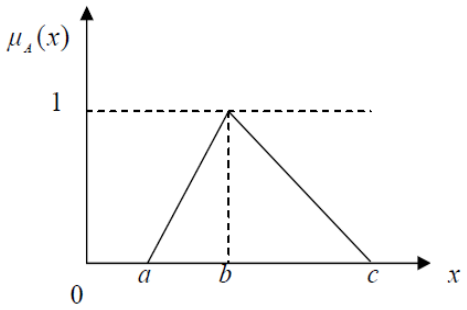

Figure 1. Triangular fuzzy membership function.

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:

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

where $s_{i,t}$ is the i-th ranked stock at time t; $R_t(\cdot)$ 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:

$$R_c = \prod_{t=1}^{n} \overline{R_t}. \quad (4)$$

C. *Model optimization*

The performance of the stock selection model is determined by the set of input features F, the set of stock sorting indicators I, the weights of the fundamental variables W, and the set of parameters θ to the fuzzy membership function. Therefore, we expect that the selection of optimal subsets of features F, and the optimization of I, W and θ will be critical to the effectiveness of the stock selection model. In this study, we propose to use Genetic Algorithms for simultaneous optimization with respect to these tasks. In the next subsections we describe the basics of the GA and the relevant optimization scheme for our stock selection model.

C.1 *Genetic algorithms*

Genetic algorithms were developed by Holland [20] and have been used as computational models of natural

evolutionary systems and as adaptive algorithms for solving optimization problems. GAs operate on an evolving population of artificial organisms, or agents. Each agent is comprised of a genotype (often a binary string) encoding a solution to some problem and a phenotype (the solution itself). GAs regularly start with a population of randomly generated agents within which solution candidates are embedded. In each iteration, a new generation is created by applying variations, such as crossover and mutation, to promising candidates selected according to probabilities biased in favor of the relatively fit agents. As a result, evolution occurs by iterated stochastic variation of genotypes, and selection of the best phenotypes in an environment according to how well the respective solution solves a problem (or problem-specific fitness function). Successive generations are created in the same manner until a well-defined termination criterion is met. The core of this class of algorithms lies in the production of new genetic structures along the course of evolution, thereby providing innovations to solutions for the problem at hand. The steps of a simple GA can be described in the following:

Step 1: Randomly generate an initial population of l agents, each being an n-bit genotype (chromosome).
Step 2: Evaluate each agent's fitness.
Step 3: Repeat until l offspring have been created.
  a) select a pair of parents for mating;
  b) apply variation operators (crossover and mutation);
Step 4: Replace the current population with the new population.
Step 5: Go to Step 2 until terminating condition.

C.2 *Chromosome encoding and fitness function*

Among many paradigms of search algorithms GAs have been proven to have an advantage over traditional optimization methods in problems with many complex, discontinuous constraints in the search space. This methodology contributes for a global, population-based search in the search space, in contrast with the kind of local, greedy search conducted by most rule-induction and decision-tree algorithms. Lower computation cost is a general advantage of local, greedy search algorithms. However, the solution quality achieved by these algorithms can be greatly degraded if there exists a considerable degree of feature interactions, which is usually the case for real-world problems. Since GAs can be designed to perform a global search for various combinations of sets of features that improve given optimization criteria, this class of algorithms are expected to cope better with feature interaction problems. In addition, it is also appealing to use GA's straightforward binary coding scheme to designate allele '1' or '0' to represent a feature being selected or not, respectively. Therefore, in this study we propose to use the GA method to search for optimal subsets of features for the stock selection model.

Apart from feature selection, three sets of free parameters, I, W and θ, are to be provided for the stock selection model. Here we employ the GA for simultaneous optimization of these tasks. In our overall encoding design, the composition of a chromosome is devised to consist of four portions - the candidate set of features F, the stock sorting indicators I, the weighs W and the fuzzy model parameters θ. In this study, the binary coding scheme is used to represent a chromosome. In Figure 2, 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 Figure 2 is the encoding of the set of parameters W. Figure 3 shows the detailed binary encoding for the weight of each individual variable, where the value of W_i (the weight for variable i) is encoded by loci b_{Wi}^1 through $b_{Wi}^{n_i}$. For the fuzzy model, the rightmost side of Figure 2 is the encoding of the set of parameters θ. For example, parameter a is encoded (in binary forms) by loci b_a^1 through $b_a^{n_a}$, as shown in Figure 4.

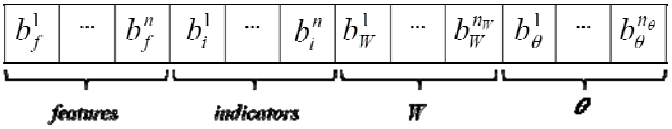

Figure 2. Chromosome encoding.

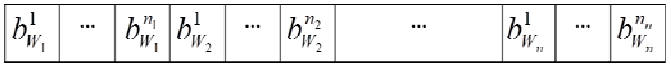
Figure 3. Detail encoding of the weighting terms W_i's.


Figure 4. Detail encoding of the fuzzy model parameters.

In our coding scheme, the portion in the chromosome representing the genotypes of parameter W_i's and θ_i'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 maxi-

mum 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, we define the fitness function of a chromosome as the annualized return of the portfolio:

$$fitness = \sqrt[n]{R_c},$$

where R_c is the cumulative total return computed by Eq. (4).

Our proposed fuzzy-GA model for stock selection is a multi-stage process, including feature selection and parameter optimization by the GA, stock scoring, score fuzzification, stock ranking and selection, as well as performance evaluation. The flowchart of this hybrid algorithm is shown in Figure 5.

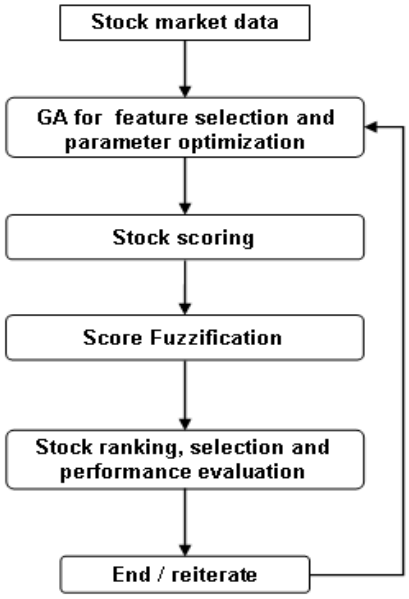

Figure 5. Flow chart of the hybrid fuzzy-GA model.

3. Data and fundamental variables

We use the constituent stocks 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 1995 to 2009. For the choice of fundamental variables, early studies indicated that several financial ratios play key roles in future stock returns. Most of them applied profitability (e.g., ROE, ROA, operating profit margin, and net profit margin), leverage (e.g., DE ratio), liquidity (e.g., cash flow ratio, current ratio and quick ratio), efficiency (e.g., inventory turnover rate and receivables turnover rate), and growth (e.g., operating rating income growth rate and net income growth rate) related ratios to examine the relationship between fundamentals and stock returns. Mukherji *et al.* [21], Jensen *et al.* [22], Danielson and Dowdell [23], Lewellen [24], Fama and French [25], and Hjalmarsson [26] also showed that the ratios relating to share price rationality, e.g., PE, PB, and PS ratios, are likely to influence future stock returns.

According to the previous literature, Table 1 provides the aforementioned six attributes that are to be employed for this study, including fifteen financial ratios. For each year, investable stocks are described by these fifteen financial ratios and their historical returns are provided. In this study, our goal is to use these attributes for the stock selection task so that a machine learning model using these attributes may be more meaningful and would offer significant insights, if any, into the real-world investment practice.

4. Empirical results

In this study, standardization was first applied to the research data — every original attribute is scaled into the range of [−1, 1] by subtracting the mean, and dividing the result by the standard deviation. This treatment is to ensure that all the attributes lie in the same parameter range, in order to prevent attributes with large ranges from overwhelming others and prediction errors may be reduced.

To examine our proposed model, stock data of all the years is first used for optimization by the GA. Stocks are ranked based on the scores obtained. Top m stocks are selected and the average yearly returns of these selected stocks are calculated. Figure 6 displays the averaged best-so-far values for annualized returns over 50 runs attained by the GA over 50 generations.[2] An averaged best-so-far performance curve is constructed by averaging the best-so-fars obtained at each generation for all 50 runs, where the vertical bars overlaying the curves represent the 95-percent confidence intervals about the means.

Figure 7 displays an illustration of the cumulative benchmark return (the product of the average yearly returns of the 200 stocks in the investment universe) and the cumulative returns of longing a number of top-ranked stocks recommended by the fuzzy-GA

[2] In order to study the change of the quality of solutions over time, a traditional performance metric for search algorithms is the "best-so-far" curve that plots the fitness of the best individual that has been seen thus far by generation n for the GA — i.e., a point in the search space that optimizes the objective function thus far. In addition, in this study the GA experiments employ a binary tournament selection [34], one-point crossover and mutation rates of 0.7 and 0.005, respectively. We also used 50 bits to represent each of the weighting terms and fuzzy model parameters.

Table 1. Attributes used in the stock selection model.

Attribute	Ratios	Description	Ref.
Share price rationality	(1) PE ratio	Price-to-earnings ratio = share price / earnings per share	[21, 23-24, 26]
	(2) PB ratio	Price-to-book ratio = share price / book value per share	[21-25]
	(3) PS Ratio	Price-to-sales ratio = share price / sales per share	[21]
Profitability	(4) ROE	Return on equity (after tax) = net income after tax / shareholders' equity	[27-28]
	(5) ROA	Return on asset (after tax) = net income after tax / total assets	[27]
	(6) OPM	Operating profit margin = operating income / net sales	[29]
	(7) NPM	Net profit margin = net income after tax / net sales	[28]
Leverage	(8) DE ratio	Debt-to-equity ratio = total liabilities / shareholders' equity	[27]
Liquidity	(9) CF ratio	Cash flow ratio = cash flow from operating activities/current liabilities	[30]
	(10) CR	Current ratio = current assets / current liabilities	[27]
	(11) QR	Quick ratio = quick assets / current liabilities	[27]
Efficiency	(12) ITR	Inventory turnover rate = cost of goods sold / average inventory	[27]
	(13) RTR	Receivables turnover rate = net credit sales / average accounts receivable	[31]
Growth	(14) OIG	Operating income growth rate = (operating income at the current year – operating income at the previous year) / operating income at the previous year	[32]
	(15) 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	[33]

model.[3] This figure shows that the portfolios of maintaining 5, 10 and 20 stocks outperform the benchmark at the end of year 2010. As a result, the optimization on the stock selection model by the GA is advantageous to stock selection.

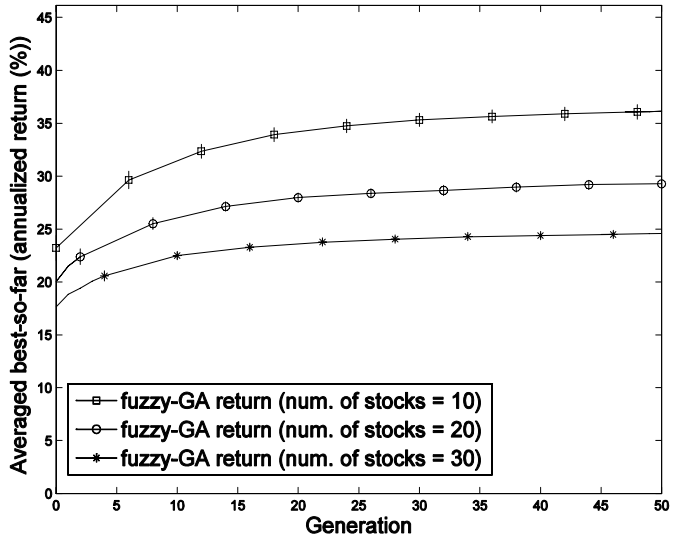

Figure 6. Best-so-far trajectories by the GA.

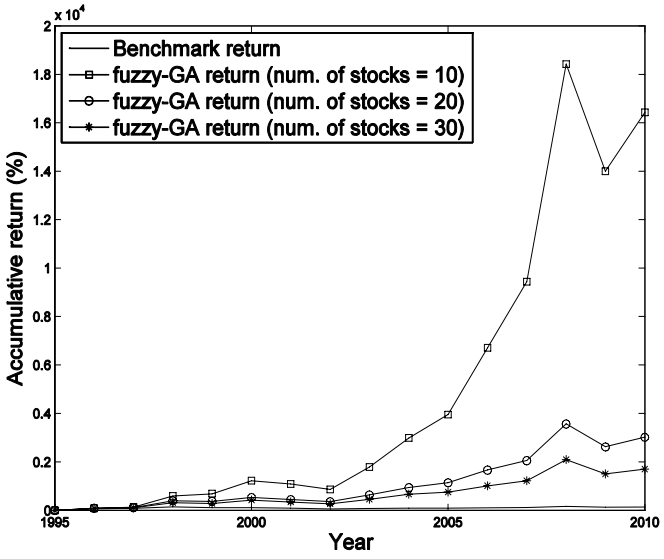

Figure 7. Cumulative returns of benchmark v.s. longing top-ranked stocks by the fuzzy-GA stock selection model.

In order to examine the validity of the methodology we proposed, further statistical validation on the model is presented next.

A. *Model validation*

In the previous subsections, the models were obtained using all the financial data available from year 1995 to 2009. In reality, the data is usually divided into two mutually exclusive parts, a training part D_r that is used to construct the models, and a testing part D_v that is used to validate the models. The goal of doing so is to examine if the models learned in the training phase are applicable to the hold-out data samples during the testing phase.

[3] Beating the benchmark has been a challenging task as discussed in [35-36]. Therefore, in this paper, we compare the performance of our stock selection model with the benchmark. Furthermore, one may also define the benchmark as the weighted return by market capitalization, but, without loss of generality, it shall suffice to examine a machine learning model by comparing its performance in accumulative returns with the product of the simple average return of all the stocks in the investment universe. In addition, "longing" is a term used in finance. In this study, "longing" a stock indicates buying it and also keeping it for another year.

Table 2. Model validation: training period (gray) and testing period (black).

N/Y	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															

In this study, we split the data into two parts: the data for the first *n* years are used to train the stock selection models and the data for the remaining years are used to validate the models. As shown in Table 2, in the first row the gray area indicates the training data (year 1995) and the model learned from this is examined by the testing data in the black area (year 1996-2009). In the second row, the training data is from year 1995 to 1996, and the testing data is from year 1997 to 2009, and so on.

Notice that this setup is different from the regular cross-validation procedure where the process of data being split into two independent sets is randomly repeated several times without taking into account the data's temporal order. However, in the stock selection study here, temporal order is critical as practically one would like to use all available data so far to train the model and hope to apply the models in the future to gain real profits.

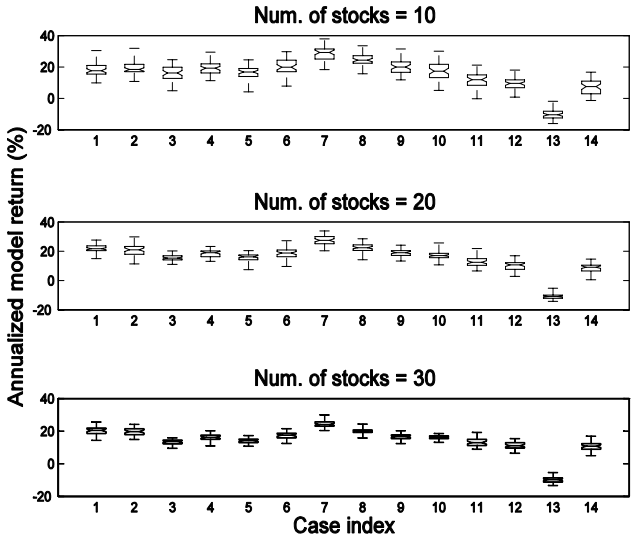

Figure 8. Box plots for the annualized model returns.

With this setup, Table 3, 4 and 5 show the model validation for selecting 10, 20 and 30 stocks by the fuzzy-GA for 50 runs. In the 10 and 20-stock model, an inspection on the means of annualized model returns shows that the model outperformed the benchmark in 13 out of 14 testing cases. For the 30-stock models, the model further outperformed the benchmark in all the 14 testing cases.

In the mean time, the variances of the annualized model returns decrease as more stocks are being selected into a portfolio. That is, selecting 30 stocks apparently yields more consistent performance on the model returns than the 10 and 20-stock cases do.

Here we also provide Figure 8 to visualize the results of Table 3, 4 and 5 through three diagrams — each of them shows a series of 14 box plots, where *x* and *y*-axis designate the case index and the annualized model return, respectively. Each box plot is generated for the corresponding 50-run fuzzy-GA model. These plots thus offer a visual gist on the spread of the model returns, and clearly selecting more stocks tends to reduce the variation in model returns.

Furthermore, it is worthwhile to investigate which features have been selected by the GA since such findings shall be important for investment in practice. Here we display in Figure 9 the ratio of the number of times a feature being selected to the total of 700 models studied.[4] As can be seen, the results are fairly consistent over the three scenarios, and the features from the share price rationality (feature 1, 2, and 3) appear relatively significant; especially feature 2 (PB ratio) has been selected most times. In addition, feature 12 (inventory turnover rate) appears relatively important across the three scenarios, yet feature 9 (cash flow ratio) also appears to be relatively important for the 20 and 30-stock scenarios.

These results thus show that the GA is able to consistently find similar subsets of significant features for the construction of the models.

Finally, we examine the accuracy and precision of our proposed model, where the accuracy and precision are defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{respectively.}$$

In these two definitions, *TP* and *TN* denote the number of true positives and true negatives, respectively. *FP* and *FN* denote the number of false positives and false negatives, respectively. In this study, a true positive occurs when the annualized model return (i.e., the annualized average return of the stocks selected by the model)

[4] We studied 50 GA models for each of the 14 cases; thus there are 700 models in total.

Table 3. Statistics of the benchmark and fuzzy-GA stock selection models for 10 stocks.

Case index	Training period	Annualized benchmark return (%)	Mean of annualized model return (%)	Variance of annualized model return (%)	Testing period	Annualized benchmark return (%)	Mean of annualized model return (%)	Variance of annualized model return (%)
1	1	-36.5079	3.328801	3.329131	2-15	4.662069	18.24054	19.56645
2	1-2	-10.1106	19.72142	4.991266	3-15	3.09983	19.1899	21.38039
3	1-3	12.71049	82.57603	7.048786	4-15	-1.45005	16.18704	19.34461
4	1-4	2.569826	59.84297	8.752561	5-15	0.749911	19.19563	16.82473
5	1-5	2.281908	59.90829	8.253837	6-15	0.711155	16.44257	16.26371
6	1-6	-1.45723	44.64023	10.08666	7-15	3.065538	19.70195	24.69397
7	1-7	-7.96379	33.01219	6.764159	8-15	10.02904	28.53181	21.17465
8	1-8	-4.53231	40.70532	4.284233	9-15	8.247265	24.64745	19.28152
9	1-9	-2.19529	43.32407	4.967174	10-15	6.59953	19.90459	19.2761
10	1-10	-1.96031	42.21486	2.86039	11-15	7.931962	17.78741	32.72187
11	1-11	-1.14807	44.02435	4.866469	12-15	8.07699	11.32397	23.75446
12	1-12	0.219419	42.75586	2.787905	13-15	5.38588	9.366132	16.47249
13	1-13	3.220311	44.49173	3.868238	14-15	-10.7904	-10.1964	9.835346
14	1-14	0.797596	37.99602	4.681656	15	7.51459	7.469343	23.85378

Table 4. Statistics of the benchmark and fuzzy-GA stock selection models for 20 stocks.

Case index	Training period	Annualized benchmark return (%)	Mean of annualized model return (%)	Variance of annualized model return (%)	Testing period	Annualized benchmark return (%)	Mean of annualized model return (%)	Variance of annualized model return (%)
1	1	-36.5079	-6.99347	1.478054	2-15	4.662069	22.05618	8.076536
2	1-2	-10.1106	12.5885	1.75532	3-15	3.09983	20.65589	15.0096
3	1-3	12.71049	66.80848	2.118662	4-15	-1.45005	15.40438	4.669493
4	1-4	2.569826	45.61831	2.230825	5-15	0.749911	18.25128	5.44494
5	1-5	2.281908	48.46778	3.193698	6-15	0.711155	15.57396	6.825322
6	1-6	-1.45723	35.58544	1.775418	7-15	3.065538	18.62363	9.692867
7	1-7	-7.96379	24.6558	2.309639	8-15	10.02904	27.31872	9.511331
8	1-8	-4.53231	31.12519	1.069392	9-15	8.247265	22.35998	10.02519
9	1-9	-2.19529	33.82199	1.094629	10-15	6.59953	18.66502	6.042211
10	1-10	-1.96031	33.69856	0.85896	11-15	7.931962	17.32362	7.964329
11	1-11	-1.14807	34.56861	1.311064	12-15	8.07699	12.89989	10.31746
12	1-12	0.219419	33.55973	1.426682	13-15	5.38588	10.01162	10.09249
13	1-13	3.220311	36.67187	1.802618	14-15	-10.7904	-10.9368	3.490574
14	1-14	0.797596	30.87962	0.943495	15	7.51459	8.602114	9.534986

does outperform the benchmark; otherwise, the model generates a false positive. Analogously, a true negative occurs when the annualized average return of the unselected stocks does underperform the benchmark; otherwise, the model generates a false negative.

As an illustration, Table 6, 7, 8 display the distributions of *TP*, *FP*, *TN* and *FN* over the 14 testing cases for the 10, 20 and 30-stock model, respectively. In Table 9, we summarize the results of the accuracy and precision calculated for the 10, 20 and 30-stock models, which appear to be very good. Thus all the results so far provide promising evidence for the effectiveness of our proposed methodology.

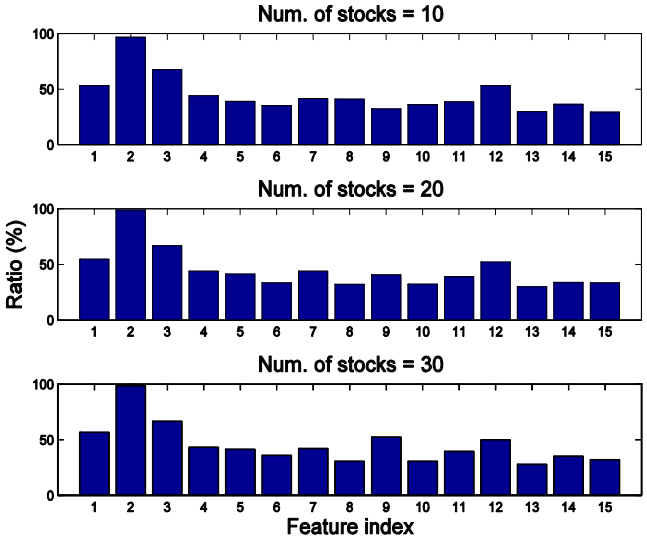

Figure 9. Ratios of features being selected.

Table 5. Statistics of the benchmark and fuzzy-GA stock selection models for 30 stocks.

Case index	Training period	Annualized benchmark return (%)	Mean of annualized model return (%)	Variance of annualized model return (%)	Testing period	Annualized benchmark return (%)	Mean of annualized model return (%)	Variance of annualized model return (%)
1	1	-36.5079	-11.7752	0.611969	2-15	4.662069	20.16443	5.565993
2	1-2	-10.1106	9.934707	1.196377	3-15	3.09983	19.54408	6.078497
3	1-3	12.71049	57.19417	1.57401	4-15	-1.45005	13.32949	2.803982
4	1-4	2.569826	38.00949	1.362032	5-15	0.749911	16.08497	3.947759
5	1-5	2.281908	41.37395	2.523244	6-15	0.711155	14.15359	2.243749
6	1-6	-1.45723	29.81098	1.524654	7-15	3.065538	17.22366	4.276182
7	1-7	-7.96379	20.73429	1.464687	8-15	10.02904	24.33025	4.754411
8	1-8	-4.53231	26.09325	1.477472	9-15	8.247265	19.83508	3.588218
9	1-9	-2.19529	28.05908	0.644863	10-15	6.59953	16.52707	3.572116
10	1-10	-1.96031	28.17533	0.581568	11-15	7.931962	16.15223	1.845652
11	1-11	-1.14807	28.28664	0.548697	12-15	8.07699	13.12556	5.389588
12	1-12	0.219419	27.81914	0.676443	13-15	5.38588	11.2536	3.962449
13	1-13	3.220311	30.89617	0.511484	14-15	-10.7904	-9.86467	2.988263
14	1-14	0.797596	25.73027	0.425362	15	7.51459	10.72187	7.686896

Table 6. Distributions of *TP*, *FP*, *TN*, *FN* for the 10-stock model.

	Case index													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TP	50	50	50	50	50	50	50	50	50	47	39	42	28	25
FP	0	0	0	0	0	0	0	0	0	3	11	8	22	25
TN	50	50	50	50	50	50	50	50	50	47	39	42	30	25
FN	0	0	0	0	0	0	0	0	0	3	11	8	20	25

Table 7. Distributions of *TP*, *FP*, *TN*, *FN* for the 20-stock model.

	Case index													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TP	50	50	50	50	50	50	50	50	50	50	48	44	21	34
FP	0	0	0	0	0	0	0	0	0	0	2	6	29	16
TN	50	50	50	50	50	50	50	50	50	50	48	45	22	34
FN	0	0	0	0	0	0	0	0	0	0	2	5	28	16

Table 8. Distributions of *TP*, *FP*, *TN*, *FN* for the 30-stock model.

	Case index													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TP	50	50	50	50	50	50	50	50	50	50	50	50	34	44
FP	0	0	0	0	0	0	0	0	0	0	0	0	16	6
TN	50	50	50	50	50	50	50	50	50	50	50	50	34	44
FN	0	0	0	0	0	0	0	0	0	0	0	0	16	6

Table 9. Accuracy and precision for the 10, 20 and 30-stock models.

	10-stock model	20-stock model	30-stock model
Accuracy	0.9029	0.9257	0.9686
Precision	0.9014	0.9243	0.9686

5. Conclusions

In this paper we presented a hybrid fuzzy-GA methodology for stock selection. Using the developed fuzzy-based scoring mechanism for a set of stocks, top-ranked stocks can be selected as components in a portfolio. In the mean time, the GA was employed for optimization of model parameters and feature selection simultaneously. In this study, we have shown that feature selection can shed light on which features play more important roles in our proposed model, which shall be important to stock selection in practice. We have also used several means to evaluate the fuzzy-GA models statistically and validated the effectiveness of this method by comparing the returns of the models with that of the benchmark. The empirical results showed that the investment returns provided by our proposed methodology can significantly outperform the benchmark. Therefore, we expect this hybrid model to advance the research in computational finance and provide a promising solution to stock selection in practice.

In the future, a plausible research direction is to employ more advanced fuzzy and scoring models to investigate how performance of portfolios can be further improved. In addition, in our current model, we consider the first several years as the training set and the next several years as the test set. This may not be sufficient to infer accurately plausible scores for the stocks in the period from 1996 till 2009 if the model uses only the stocks of 1995 for training. Therefore, in the future work, we intend to study a model that would be capable of generating more time-dependent patterns to account for the impact of the most recent history of the stocks. Finally, because investment return and risk management appear

to be two distinct objectives, in the future work, we expect that a study for simultaneous optimization on these multi-objectives is also a promising research area.

Acknowledgment

This work is fully supported by the National Science Council, Taiwan, Republic of China, under grant number NSC 99-2221-E-390-032.

References

[1] A. Mochón, D. Quintana, and Y. Sáez, "Soft computing techniques applied to finance," *Applied Intelligence*, vol. 29, no. 2, pp.111-115, 2008.

[2] D. Zhang and L. Zhou, "Discovering golden nuggets: data mining in financial application," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Review*, vol. 34, no. 4, pp. 513-522, 2004.

[3] K. K. Lai, L. Yu, S. Wang, and C. Zhou, "A double-stage genetic optimization algorithm for portfolio selection," In *Proc. of the 13th International Conference on Neural Information Processing*, pp. 928-937, 2006.

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

[5] N. Chapados and 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.

[6] K. J. Kim and 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.

[7] M. Caplan and Y. Becker, "Lessons Learned Using Genetic Programming in A Stock Picking Context," in: U.-M. O'Reilly, T. Yu, R. Riolo, B. Worzel (Eds.), *Genetic Programming Theory and Practice II*, Springer, Chapter 6, pp. 87-102, Ann Arbor, Michigan, 2004.

[8] Y. Becker, P. Fei, and 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*, Springer, Chapter 12, pp. 315-334, Ann Arbor, Michigan, 2006.

[9] T. C. Chu, C. T. Tsao, and Y. R. Shiue, "Application of fuzzy multiple attribute decision making on company analysis for stock selection," In *Proc. of Soft Computing on Intelligent Systems and Information Processing*, pp. 509-514, 1996.

[10] M. R. Zargham and M. R. Sayeh, "A web-based information system for stock selection and evaluation," In *Proc. of the First International Workshop on Advance Issues of E-Commerce and Web-Based Information Systems*, pp. 81-83, 1999.

[11] B. R. Chang, "Applying nonlinear generalized autoregressive conditional heteroscedasticity to compensate ANFIS outputs tuned by adaptive support vector regression," *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1832-1850, 2006.

[12] B. R. Chang, "Resolving the forecasting problems of overshoot and volatility clustering using ANFIS coupling nonlinear heteroscedasticity with quantum tuning," *Fuzzy Sets and Systems*, vol. 159, no. 23, pp. 3183-3200, 2008.

[13] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems*, vol. 13, no. 2, pp. 44-49, 1998.

[14] G. John, R. Kohavi, and K. Peger, "Irrelevant features and the subset selection problem," In *Proc. of the 11th International Conference on Machine Learning*, pp. 121-129, 1994.

[15] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no.1-2, pp. 273-324, 1997.

[16] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert Systems with Applications*, vol. 31, pp. 231-240, 2006.

[17] S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," Department of Information and Computer Science, University of California, Irvine, CA, 1998, http://archive.ics.uci.edu/ml/.

[18] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[19] J. Piotroski, "Value investing: The use of historical financial statement information to separate winners from losers," *Journal of Accounting Research*, vol. 38, pp. 1-41, 2000.

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

[21] S. Mukherji, M. S. Dhatt, and Y. H. Kim, "A fundamental analysis of Korean stock returns," *Financial Analysts Journal*, vol. 53, no. 3, pp. 75-80, 1997.

[22] G. R. Jensen, R. R. Johnson, and J. M. Mercer, "New evidence on size and price-to-book effects in stock returns," *Financial Analysts Journal*, vol. 53, no. 6, pp. 34-42, 1997.

[23] M. G. Danielson and T. D. Dowdell, "The re-

turn-stages valuation model and the expectations within a firm's P/B and P/E ratios," *Financial Management*, vol. 30, no. 2, pp. 93-124, 2001.
[24] J. Lewellen, "Predicting returns with financial ratio," *Journal of Financial Economics*, vol. 74, no. 2, pp. 209-235, 2004.
[25] E. F. Fama and K. R. French, "Average returns, B/M, and share issue," *Journal of Finance*, vol. 63, no. 6, pp. 2971-2995, 2008.
[26] E. Hjalmarsson, "Predicting global stock returns," *Journal of Financial and Quantitative Analysis*, vol. 45, no. 1, pp. 49-80, 2010.
[27] 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, no. 2, pp. 84-102, 2004.
[28] R. Bauer, N. Guenster, and R. Otten, "Empirical evidence on corporate governance in Europe: the effect on stock returns, firm value and performance," *Journal of Asset Management*, vol. 5, no. 2, pp. 91-104, 2004.
[29] M. T. Soliman, "The use of DuPont analysis by market participants," *The Accounting Review*, vol. 83, no. 3, pp. 823-853, 2008.
[30] W. E. Ferson and C. R. Harvey, "Fundamental determinants of national equity market returns: A perspective on conditional asset pricing," *Journal of Banking & Finance*, vol. 21, pp. 1625-1665, 1998.
[31] 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, no. 3, pp. 99-116, 2006.
[32] D. Ikenberry and J. Lakonishok, "Corporate governance through the proxy contest: evidence and implications," *Journal of Business*, vol. 66, no. 3, pp. 405-435, 1993.
[33] G. Sadka and R. Sadka, "Predictability and the earnings-returns relation," *Journal of Financial Economics*, vol. 94, no. 1, pp. 87-93, 2009.
[34] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," *Foundation of Genetic Algorithms*, pp. 69-93, 1991.
[35] B. A. Sensoy, "Performance evaluation and self-designated benchmark indexes in the mutual fund industry," *Journal of Financial Economics*, vol. 92, no. 1, pp. 25-39, 2009.
[36] R. Shukla and S. Singh, "A performance evaluation of global equity mutual funds: Evidence from 1988-95," *Global Finance Journal*, vol. 8, no. 2, pp. 279-293, 1997.

Dr. Chien-Feng Huang is currently Assistant Professor of the Department of Computer Science and Information Engineering at National University of Kaohsiung in Kaohsiung City, Taiwan. In 2002, he earned his doctoral degree from the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, USA. He has published more than 50 scientific papers on various journals, conference proceedings and books. His current research interests include computational finance, data mining, intelligent computing and data compression.

Dr. Bao Rong Chang is currently a Full Professor in the Department of Computer Science and Information Engineering at National University of Kaohsiung in Kaohsiung City, Taiwan. He is currently conducting a cloud computing program in the same university and leading a cloud computing group for executing several projects granted from both National Science Council and Ministry of Education in Taiwan. In 1990, he earned his ME degree from the Department of Electrical and Computer Engineering, University of Missouri-Columbia, USA, and his Ph.D. in 1994 at the same University. His current research interests include Cloud Computing, Embedded System, Multimedia Applications, and Intelligent Computation.

Mr. Dun-Wei Cheng is currently pursuing the Ph.D. degree in the Department of Computer Science and Information Engineering at National Cheng Kung University in Tainan City, Taiwan. He received the M.S. degree from the Department of Computer Science and Information Engineering at National University of Kaohsiung, in 2011. His current research interests include financial computing, artificial intelligence and computer vision.

Dr. Chih-Hsiang Chang is Associate Professor of the Department of Finance at National University of Kaohsiung, Taiwan, ROC. His main research interests are behavioral finance, investments, market microstructure, and corporate finance. He has published more than 35 articles in various journals, including Academia Economic Papers, Accounting and Finance, Asia-Pacific Journal of Financial Studies, Asia Pacific Management Review, Canadian Journal of Administrative Sciences, China & World Economy, Chinese Economy, European Financial Management, International Journal of Fuzzy Systems, International Review of Economics & Finance, Journal of Financial Studies, Journal of Management & Systems, Management Review, Managerial Finance, NTU Management Review, Pacific-Basin Finance Journal, Sun Yat-sen Management Review, among others.