

A Study of a Hybrid Evolutionary Fuzzy Model for Stock Selection

Chien-Feng Huang
Department of Computer
Science and Information
Engineering
National University of
Kaohsiung
Kaohsiung, Taiwan, R.O.C.
cfhuang15@nuk.edu.tw

Chih-Hsiang Chang
Department of Finance
National University of
Kaohsiung
Kaohsiung, Taiwan, R.O.C.
cch@nuk.edu.tw

Bao Rong Chang
Department of Computer
Science and Information
Engineering
National University of
Kaohsiung
Kaohsiung, Taiwan, R.O.C.
brchang@nuk.edu.tw

Dun-Wei Cheng
Department of Computer
Science and Information
Engineering
National University of
Kaohsiung
Kaohsiung, Taiwan, R.O.C.
m0985504@mail.nuk.edu.tw

Abstract—Stock selection has long been 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 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. Top-ranked stocks can thus be selected to form a portfolio. Furthermore, we employ GA for optimization of model parameters and feature selection for input variables to the stock scoring model. We will show that the investment returns provided by our proposed methodology significantly outperform the benchmark return. Based upon the promising results obtained, we expect this hybrid fuzzy-GA methodology to advance the research in soft computing for finance and provide an effective solution to stock selection in practice.

Keywords- Stock selection; fuzzy models; genetic algorithms; model validation

I. 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.

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 optimization of parameters, 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 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.

II. 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 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.

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-asset (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:

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

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

In addition, let W_j denote the weight of the j -th variable. Then the total score of stock i at time t , $S_{i,t}$, 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 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 Eq. (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 , $y_{i,t}(\mathbf{W}, \boldsymbol{\theta}) \in R$, is defined as

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

where $\mathbf{W}$ denotes the vector of the weights of the input features (i.e., the fundamental variables) used by the stock scoring model and $\boldsymbol{\theta}$ 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, Fig. 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.

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

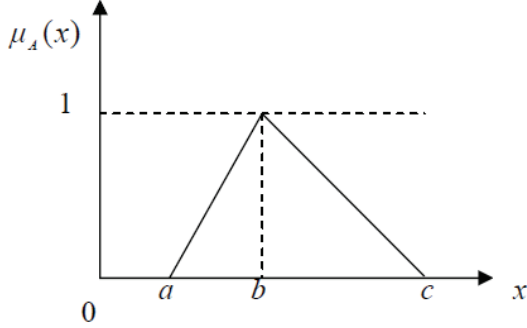


Figure 1. Triangular fuzzy membership function

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:

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

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:

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

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 $\bar{R}_t$ 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, $\bar{R}_t$, of the stocks in a portfolio over n consecutive years as:

$$R_c = \prod_{t=1}^n \bar{R}_t. \quad (4)$$

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 $\boldsymbol{\theta}$ to the fuzzy membership function. Therefore, we expect that the selection of optimal subsets of features F , and the optimization of I , W and $\boldsymbol{\theta}$ 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 are shown 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. Therefore, 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 Fig. 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 Fig. 2 is the encoding of the set of parameters W . Fig. 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 $b_{W_i}^1$ through $b_{W_i}^{n_i}$. For the fuzzy model, the rightmost side of Fig. 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 Fig. 4.

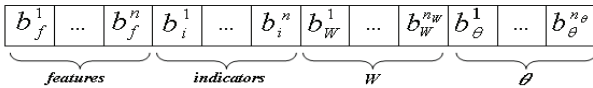


Figure 2. Chromosome encoding

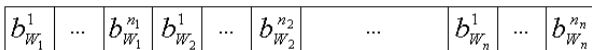


Figure 3. Detail encoding of the weighting terms W_i 's

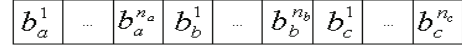


Figure 4. Detail encoding of the fuzzy model parameters

In our coding scheme, the portion in the chromosome representing the genotypes of parameter W_i 's and θ '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:

$$y = \min_y + \frac{d}{2^l - 1} \times (\max_y - \min_y), \quad (5)$$

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

$$fitness = \sqrt[n]{R_c}, \quad (6)$$

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 Fig. 5.

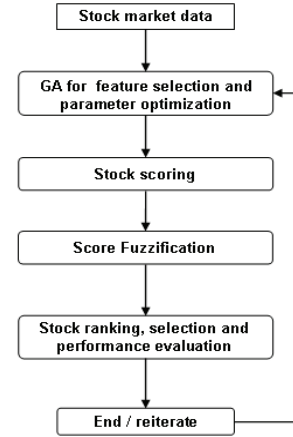


Figure 5. Flow chart of the hybrid fuzzy-GA model

III. DATA AND FUNDAMENTAL VARIABLES

We use the constituent stocks of the 50 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

Table I. Attributes used in the stock selection model

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

(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., DB ratio), liquidity (e.g., 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 fourteen financial ratios. For each year, investable stocks are described by these fourteen financial ratios and their historical returns are provided.

IV. 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

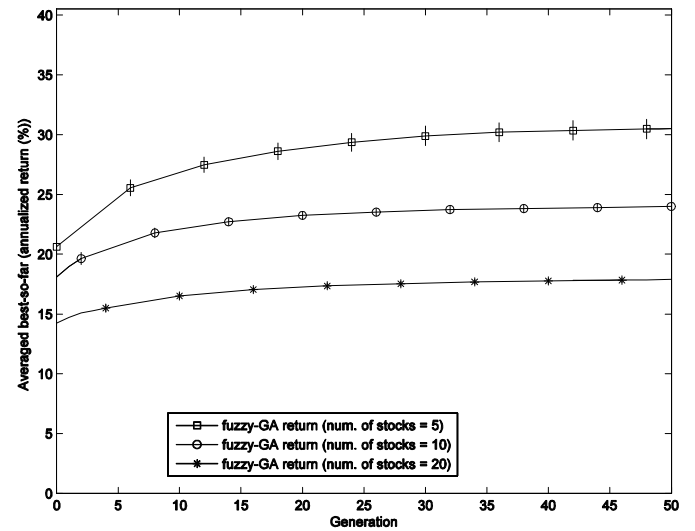


Figure 6. Best-so-far curves by the GA

the average yearly return of these selected stocks are calculated. Fig. 6 displays the averaged best-so-far values for annualized returns over 50 runs attained by the GA over 50 generations.¹ The averaged best-so-far performance curve is calculated by averaging the best-so-fars obtained at each generation for all 50 runs, where the vertical bars overlaying

¹ 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 [33], one-point crossover and mutation rates of 0.7 and 0.005, respectively.

the curves represent the 95-percent confidence intervals about the means.

Fig. 7 displays an illustration of the cumulative benchmark return (the product of the average yearly returns of the 50 stocks in the investment universe) and the cumulative returns of longing a number of top-ranked stocks recommended by the fuzzy-GA model.² This figure shows that the portfolios of maintaining 5, 10 and 20 stocks outperform the benchmark at the end of year 2009. As a result, the optimization on the stock selection model by the GA is advantageous to stock selection.

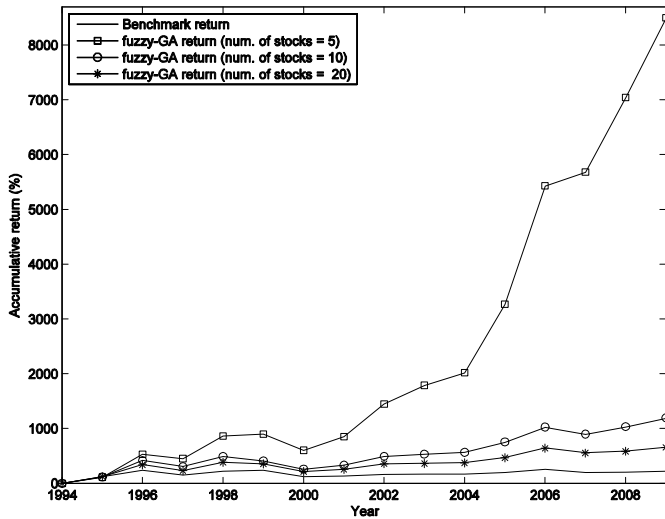


Figure 7. Cumulative returns of benchmark v.s. longing top-ranked stocks by the fuzzy-GA stock selection model

In order to examine the validity of the methodology we proposed, further statistical validation on the model is conducted. In reality, the learned model has to be tested by

unseen data. Here, we use the stock data of the first several years to train the fuzzy-GA model, and the remaining years are used for testing the learned model. Thus the average yearly return of the selected stocks for each of the testing year is calculated. The average yearly returns are then compounded to obtain the cumulative total return of the portfolio over the testing years.

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 to apply the models in the future to gain profits.

Here we compare the cumulative benchmark return (the product of the average yearly returns of the 50 stocks in the investment universe) and the cumulative average return of longing a number of top-ranked stocks (5 and 20 stocks) using our stock selection model. Table 2 and 3 show the model validation for selecting 5 and 20 stocks by the fuzzy-GA model for 50 runs. In the 5-stock model, an inspection on the means of annualized model returns shows that 13 out of 14 for the testing cases outperform the annualized benchmark return. For the 20-stock model, in all the 14 testing cases, the means of annualized model returns all outperform the annualized benchmark return.

In the mean time, the standard deviations and standard errors of the annualized model returns decrease as more stocks are being selected into a portfolio. That is, selecting 20 stocks apparently yields more consistent performance on the model

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

Training period	Annualized benchmark return	Mean of annualized model return	Standard Deviation	Standard error	Testing period	Annualized benchmark return	Mean of annualized model return	Standard Deviation	Standard error
1995	23.1487	61.4170	2.61412	0.36969	1996-2009	4.25672	1.1731	5.34558	0.75598
1995-1996	53.9684	162.6117	10.33728	1.46191	1997-2009	-0.54716	-0.5062	4.02765	0.56959
1995-1997	15.3332	76.9770	3.14302	0.44449	1998-2009	3.07863	5.2559	5.66329	0.80091
1995-1998	22.4658	87.1291	5.02817	0.71109	1999-2009	-0.17093	0.1383	5.11039	0.72271
1995-1999	19.1951	69.1519	5.39052	0.76233	2000-2009	-0.85761	0.9131	5.79058	0.81891
1995-2000	3.7129	43.7078	3.15297	0.44589	2001-2009	6.57475	10.2523	7.24027	1.02393
1995-2001	4.0824	40.2833	2.63405	0.37251	2002-2009	6.60575	13.3479	4.55598	0.64431
1995-2002	6.5905	41.9213	2.32244	0.32844	2003-2009	4.09937	10.1117	5.37307	0.75986
1995-2003	6.1517	39.6470	2.09784	0.29668	2004-2009	4.33343	7.6255	7.19403	1.01739
1995-2004	5.5582	36.0228	1.25193	0.17705	2005-2009	5.14611	11.0043	8.36337	1.18275
1995-2005	6.5087	35.5869	1.82544	0.25815	2006-2009	2.48545	9.02378	7.69611	1.08839
1995-2006	8.124	36.9053	1.67149	0.23638	2007-2009	-4.73667	-1.64134	9.27202	1.31126
1995-2007	5.5105	33.5921	1.64090	0.23205	2008-2009	4.83837	5.25787	9.33182	1.31972
1995-2008	5.3590	31.5289	1.71741	0.24287	2009	6.28742	15.3057	12.88159	1.82173

² Beating the benchmark has been a challenging task as discussed in [34-35]; therefore, in this paper, we compare the performance of our stock selection model with the benchmark.

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

Training period	Annualized benchmark return	Mean of annualized model return	Standard Deviation	Standard error	Testing period	Annualized benchmark return	Mean of annualized model return	Standard Deviation	Standard error
1995	23.1487	39.7684	1.57912	0.2233	1996-2009	4.25672	4.49396	3.07153	0.43438
1995-1996	53.9684	97.5861	2.23359	0.3158	1997-2009	-0.54716	2.12094	2.45705	0.34748
1995-1997	15.3332	39.8031	1.23430	0.1745	1998-2009	3.07863	7.35682	2.40687	0.34038
1995-1998	22.4658	50.1824	1.84520	0.2609	1999-2009	-0.17093	1.37174	2.33844	0.33070
1995-1999	19.1951	40.3479	1.53589	0.2172	2000-2009	-0.85761	0.74860	2.66413	0.37676
1995-2000	3.7129	22.0304	1.36951	0.1936	2001-2009	6.57475	11.0353	1.74344	0.24656
1995-2001	4.0824	21.3027	1.12551	0.1591	2002-2009	6.60575	11.2566	1.96797	0.27831
1995-2002	6.5905	24.0367	1.06560	0.1507	2003-2009	4.09937	8.04446	2.02825	0.28683
1995-2003	6.1517	22.5926	0.96115	0.1359	2004-2009	4.33343	7.75598	2.02601	0.28652
1995-2004	5.5582	20.6871	0.81464	0.1152	2005-2009	5.14611	9.91133	2.24343	0.31727
1995-2005	6.5087	20.3811	0.80881	0.1143	2006-2009	2.48545	8.27924	2.47071	0.34941
1995-2006	8.1249	22.7893	0.79199	0.1120	2007-2009	-4.73667	-1.34713	2.26997	0.32102
1995-2007	5.5105	19.4647	0.53036	0.0750	2008-2009	4.83837	6.23260	2.77440	0.39236
1995-2008	5.3590	18.6817	0.71449	0.1010	2009	6.28742	7.22983	4.71266	0.66647

return from various runs than the 5-stock cases do. As a result, these validations provide a statistical evidence that our proposed stock selection methodology is promising for investment in research and practice.

V. CONCLUSIONS

In this paper we presented a hybrid fuzzy-GA methodology for stock selection. Based on the designed 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 for the model. We have evaluated the fuzzy-GA models statistically and validated the effectiveness of this method by comparing the investment return of the models with that of the benchmark. The empirical results showed that the investment return provided by our proposed methodology can significantly outperform the benchmark return. 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 generate a feasible model; e.g., it might be difficult to infer plausible scores for 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 more recent past 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, Y. Sáez, "Soft computing techniques applied to finance", *Applied Intelligence*, vol.29(2), 2008, pp.111-115.
- [2] D. Zhang, L. Zhou, "Discovering golden nuggets: data mining in financial application", *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Review*, vol.34(4), 2004, pp. 513–522.
- [3] K. K. Lai, L. Yu, S. Wang, C. Zhou, "A double-stage genetic optimization algorithm for portfolio selection", the 13th International Conference on Neural Information Processing, 2006, pp. 928–937.
- [4] T. S. Quah, B. Srinivasan, "Improving returns on stock investment through neural network selection", *Expert Systems with Applications*, vol.17, 1999, pp. 295–301.
- [5] N. Chapados, Y. Bengio, "Cost functions and model combination for VaR-based asset allocation using neural networks", *IEEE Transactions on Neural Networks*, vol. 12, 2001, pp. 890–906.
- [6] 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, 2000, pp. 125–132.
- [7] M. Caplan, 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, Ann Arbor, Michigan, 2004, pp. 87–102, Chapter 6.
- [8] 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, 2006, pp. 315–334, Chapter 12.
- [9] T. C. Chu, C. T. Tsao, Y. R. Shiue, "Application of fuzzy multiple attribute decision making on company analysis for stock selection", in:

- Proceedings of Soft Computing on Intelligent Systems and Information Processing, 1996, pp. 509–514.
- [10] M. R. Zargham, M. R. Sayeh, “A web-based information system for stock selection and evaluation”, in: Proceedings of the First International Workshop on Advance Issues of E-Commerce and Web-Based Information Systems, 1999, pp. 81-83.
 - [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(13), 2006, pp. 1832-1850.
 - [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(23), 2008, pp. 3183-3200.
 - [13] J. Yang, V. Honavar, “Feature subset selection using a genetic algorithm”, *IEEE Intelligent Systems*, vol.13(2), 1998, pp. 44-49.
 - [14] G. John, R. Kohavi, K. Peger, “Irrelevant features and the subset selection problem,” in: Proceedings of the 11th international conference on machine learning, 1994, pp. 121–129.
 - [15] R. Kohavi, G. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97(1–2), 1997, pp. 273–324.
 - [16] C. L. Huang, C. J. Wang, “A GA-based feature selection and parameters optimization for support vector machines,” *Expert Systems with Applications*, vol. 31, 2006, p. 231–240.
 - [17] S. Hettich, C. L. Blake, C. J. Merz, “UCI repository of machine learning databases”, Department of Information and Computer Science, University of California, Irvine, CA. <http://archive.ics.uci.edu/ml/>.
 - [18] I. H. Witten, 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, 2000, pp. 1–41.
 - [20] J. H. Holland, “Adaptation in natural and artificial systems,” University of Michigan Press, Ann Arbor, Michigan, 1975.
 - [21] S. Mukherji, M. S. Dhatt, Y. H. Kim, “A fundamental analysis of Korean stock returns,” *Financial Analysts Journal*, vol.53(3), 1997, pp. 75-80.
 - [22] G. R. Jensen, R. R. Johnson, J. M. Mercer, “New evidence on size and price-to-book effects in stock returns,” *Financial Analysts journal*, vol.53(6), 1997, pp.34-42.
 - [23] M. G. Danielson, T. D. Dowdell, “The return-stages valuation model and the expectations within a firm’s P/B and P/E ratios,” *Financial Management*, vol. 30(2), 2001, pp.93-124.
 - [24] J. Lewellen, “Predicting returns with financial ratio,” *Journal of Financial Economics*, vol. 74(2), 2004, pp. 209-235.
 - [25] E. F. Fama, K. R. French, “Average returns, B/M, and share issue,” *Journal of Finance*, vol. 63(6), 2008, pp. 2971-2995.
 - [26] E. Hjalmarrsson, “Predicting global stock returns,” *Journal of Financial and Quantitative Analysis*, vol. 45(1), 2010, pp. 49-80.
 - [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(2), 2004, 84-102.
 - [28] 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), 2004, 91-104.
 - [29] M. T. Soliman, “The use of DuPont analysis by market Participants,” *The Accounting Review*, vol. 83(3), 2008, 823-853.
 - [30] 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), 2006, 99-116.
 - [31] D. Ikenberry, J. Lakonishok, “Corporate governance through the proxy contest: evidence and implications,” *Journal of Business*, vol. 66(3), 1993, 405-435.
 - [32] G. Sadka, R. Sadka, “Predictability and the earnings-returns relation,” *Journal of Financial Economics*, vol. 94(1), 2009, 87-93.
 - [33] D. E. Goldberg, K. Deb, “A comparative analysis of selection schemes used in genetic algorithms,” *Foundation of Genetic Algorithms*, 1991, pp. 69-93.
 - [34] B. A. Sensoy, “Performance Evaluation and Self-Designated Benchmark Indexes in the Mutual Fund Industry,” *Journal of Financial Economics*, vol. 92(1), 2009, 25-39.
 - [35] R. Shukla, S. Singh, “A Performance Evaluation of Global Equity Mutual Funds: Evidence from 1988-95,” *Global Finance Journal*, vol. 8(2), 1997, 279-293.