Prediction of the price for Stock Index Futures using Integrated Artificial Intelligence Techniques with Categorical Preprocessing

Kyoung-jae Kim and Ingoo Han

Graduate School of Management
Korea Advanced Institute of Science and Technology
207-43 Cheongryanri-Dong Dongdaemun-Gu, Seoul, Korea
Telephone: 82-2-958-3663, Fax: 82-2-958-3604, E-mail: ingoohan@msd.kaist.ac.kr

Abstract

Previous studies in stock market predictions using artificial intelligence techniques such as artificial neural networks and case-based reasoning, have focused mainly on spot market prediction. Korea launched trading in index futures market (KOSPI 200) on May 3, 1996, then more people became attracted to this market. Thus, this research intends to predict the daily up/down fluctuant direction of the price for KOSPI 200 index futures to meet this recent surge of interest. The forecasting methodologies employed in this research are the integration of genetic algorithm and artificial neural network (GAANN) and the integration of genetic algorithm and case-based reasoning (GACBR). Genetic algorithm was mainly used to select relevant input variables.

This study adopts the categorical data preprocessing based on expert's knowledge as well as traditional data preprocessing. The experimental results of each forecasting method with each data preprocessing method are compared and statistically tested.

Artificial neural network and case-based reasoning methods with best performance are integrated. Out-of-the-Model Integration and In-Model Integration are presented as the integration methodology.

The research outcomes are as follows;

First, genetic algorithms are useful and effective method to select input variables for AI techniques. Second, the results of the experiment with categorical data preprocessing significantly outperform that with traditional data preprocessing in forecasting up/down fluctuant direction of index futures price. Third, the integration of genetic algorithm and case-based reasoning (GACBR) outperforms the integration of genetic algorithm and artificial neural network (GAANN). Forth, the integration of genetic algorithm, case-based reasoning and artificial neural network (GAANN-GACBR, GACBRNN and GANNCBR) provide worse results than GACBR.

Key words: the price for Stock Index Futures, Genetic Algorithm, Artificial Neural Network, Case-Based Reasoning, Integrated Model, Categorical Preprocessing

Introduction

It is known that stock market shows a nonlinear pattern. In this aspect, there may be some limitations to analyze the stock market with linear models. Therefore, artificial intelligence (AI) techniques are often used for the nonlinear pattern analysis such as stock market analysis. Previous studies in stock market predictions using AI techniques such as artificial neural networks and case-based reasoning, have focused mainly on spot market prediction. Korea launched trading in index futures market (KOSPI 200) on May 3, 1996, then more people became attracted to this market, because the investor can avoid or reduce the risk of stock investment via stock index futures. The lack of the research on the index futures market in Korea has not met the people's interest. Thus, this research intends to predict the daily up/down fluctuant direction of the price for KOSPI 200 index futures to meet this recent surge of interest.

This research has three objectives as follows;

First, this study is to test the predictability of the price for stock index futures by predicting KOSPI 200 futures price.

Second, this study is to analyze the difference of prediction accuracy using different data preprocessing method.

Third, the prediction accuracy of various integrated artificial intelligence techniques.

Genetic Algorithms

Genetic Algorithms (GAs) are search algorithms based on the mechanics of natural selection and genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search [2].

The general procedure of GA is as follows;

First, it transforms a problem into the string, which is the structure of the chromosomes. Then, it composite population of strings and evaluate population of strings by the objective function and discards those with low evaluation values. In this way, it takes population of strings with good evaluation values and produces new offspring chromosomes by genetic operations such as reproduction, crossover, mutation etc. Finally, it repeats this procedure until evaluation values of population goes bad.

Artificial Neural Networks

Artificial Neural Networks (ANNs) are information processing systems to produce output values according to certain logic. It has certain performance characteristics in

common with biological neural networks. There are many applications of neural networks, for instance, pattern recognition, signal processing, speech production, speech recognition and applications of business such as stock market prediction, bankruptcy prediction, bond rating, credit evaluation etc.

The general procedure of Artificial Neural Network is as follows;

First, it computes of the net weighted input, and converts the net input signal to an activation level. Then transfers the net weighted sum and computes the competition function of the output function. Then computes the current error and error function, estimates the backpropagated error value. Finally it learns and adapts [10].

Case-Based Reasoning

Case-Based Reasoning (CBR) can adapt old solutions to meet new demands, use old cases to explain new situations, and critique new solutions, reason from precedents to interpret a new situation (much as lawyers do), and create an equitable solution to a new problem (much as labor mediators do) [7].

In proceeding Case-Based Reasoning, it is gave attention to following factors.

(1) Case Representation: Cases are composed of problem, description of situation, solution and results and must be represented in a structured manner.

(2) Case Indexing: Case indexing involves assigning indices to cases to facilitate their retrieval.

(3) Case Retrieval: A retrieval algorithm must retrieve the most similar cases to the new problem. Generally, nearest-neighbor method, inductive retrieval method, knowledge-based retrieval method, and combined form of other three methods are used as case retrieval [1].

(4) Case Adaptation: Once the retrieval process is executed, CBR should adapt the solution stored in the retrieved case.

Following is the general procedure of CBR with nearest-neighbor retrieval method [5].

1. Begin with current case $x(t)$.
2. Seek the L neighboring cases $x(t_i)$ in the past which are closest to $x(t)$ according to the Euclidean Distance Function:
$$d_i = \left[\sum \left| x(t_i) - x(t) \right|^M \right]^{1/M}$$
3. Compute the sum of distances:
$$d_{Tot} = \sum_{i=1}^{L} d_i$$
4. Determine the relative weight of i^{th} neighbor:
$$W_i = \frac{1}{L-1}\left[1 - \frac{d_i}{d_{Tot}}\right]$$
5. Find the successor $x(t_i + 1)$ of each case $x(t_i)$ in the set of neighbors.
6. Calculate the forecast for $t+1$ as the weighted sum of successors:
$$v(t+1) = \sum_{i=1}^{L} W_i x(t_i + 1)$$

Futures Market Analysis

Futures are the standard forms that decide the quantity and price in the certified market (trading place) at certain future point of time (delivery date). General functions of futures market are supplying information about future price of commodities, function of speculation and hedging [6].

Being different from the spot market, futures market does not have continuity of price data. That is because futures market has price data by contract. So, in futures market analysis, nearest contract data method is mainly used and incorporated in this research.

Many previous stock market analyses have used technical or fundamental indicator. In general, fundamental indicators are mostly used for long-term trend analysis while technical indicators are for short-term pattern analysis. In a study done by Van Eyden, it was found that 95 percent of the analysts in the Republic of South Africa used technical analysis. [9] In this research, we use the technical indicators as input variables.

Research Data and Experiments

(1) Research data

Research data used in this research are KOSPI 200 Futures Price. Initial available data are technical indicators such as PVI (Positive Volume Index), Stochastic%K, %D, Slow%D, Basis, Open Interest, Momentum, ROC (Rate of Change), LW%R (Larry William's%R), A/D Oscillator (Accumulation / Distribution Oscillator), ADL (Accumulation / Distribution Line), Disparity 5days, CCI (Commodity Channel Index), OSCP (Price Oscillator) and RSI (Relative Strength Index) [3].

Previous researches usually used the statistical method such as correlation test, factor analysis, stepwise regression analysis etc. This study uses genetic algorithm for input variable selection.

Genetic algorithms are performed by NeuralWorks Predict (NeuralWare, Inc.). The crossover probabilities and probability of mutation assumed to be 0.7 and 0.05 respectively.

This study incorporated two competitive methods for data preprocessing. The first is the linear scaling, which is a traditional preprocessing method. The second is the linear scaling after categorical classification. [3] The categorical classification criteria are expert knowledge-based. For example, market technicians usually regard below 25 of stochastic %k level as the signal of bear market, and above 75 as the signal of bull market and between 25 and 75 as the signal of neutral market. Therefore, it is gave a name as expert's knowledge-based preprocessing or categorical preprocessing. Table 1 shows

categorical classification criteria.

Category / Indicator	A	B	C
Stochastic% K	Below 25	25-75	above 75
Momentum		(-)	0 or (+)
CCI		(-)	0 or (+)
OSCP		(-)	0 or (+)
PVI		below MA5 of PVI	above MA5 of PVI
Stochastic Slow%D	below 25	25-75	above 75
RSI	below 30	30-70	above 70
ROC		below 100	0 or above 100
A/D Oscillator		below 0.5	0 or above 0.5

<Table 1> Categorical Classification Criteria

We assigned -1, 0 and 1 to each category A, B and C respectively.

In addition, we use two kinds of AI techniques; one is ANN and the other is CBR.

Consequently, final input variables by each preprocessing method and the kind of AI techniques are as follows. GA selects these variable sets, because each set reveals the best evaluation values respectively. GAANN means the integration of genetic algorithm and artificial neural network and GACBR means the integration of genetic algorithm and case-based reasoning in table 2.

	Traditional Preprocessing	Categorical Preprocessing
GA ANN	Stochastic%D, LW%R, Disparity5, CCI, OSCP	Stochastic%K, Momentum, CCI, OSCP, PVI
GA CBR	RSI, Momentum, ROC, LW%R, CCI, OSCP	Stochastic Slow%D, RSI, ROC, A/D Oscillator, PVI

<Table 2> Final Input Variables

(2) Experiments

Phase I: Comparison of stand-alone model

In Phase I, the difference of the hit ratio between AI techniques and between preprocessing methods are compared. Experiments are implemented by NeuroShell II 3.0 (Ward System Group Inc.) for ANN and Turbo C based program for CBR. The backpropagation algorithm and sigmoid function are used respectively in ANN. Learning rate and momentum is 0.1 respectively and initial weight is 0.3. The 10% of data for testing, 20% for holdout, and 70% for training mutually and exclusively used in order to avoid overfitting. And only 50,000 learning events since minimum average error of test set are permitted.

In experiments for CBR, we use nearest composite neighbor method. This method takes account of a "virtual" or composite neighbor whose parameters are defined by some weighted combination of actual neighbors in the case base. The key to the composite approach lies in the determination of the most effective set of weights to use in order to construct the virtual neighbor [4].

We use cross-validation method to solve insufficiency problem in the number of data and to generalize the experimental results. For given method and sample size, n, a classifier is generated using $(n-1)$ cases and tested on the single remaining case. This is repeated n times. Thus, each case in the sample is used as a test case, and each time nearly all the cases are used to design a classifier. The cross-validation error rate estimator is an almost unbiased estimator of the true error rate of a classifier [11]. In this study, first we classifying the mutually exclusive five sets which is composed of 30 of all 150 data respectively. Then it uses four sets for training and testing and the other one set for holdout. We repeated this procedure five times. Finally, we compose data set of 600 data for training and testing and 150 data for validating.

For the reason of fair comparison, we compare three sets, according to the method of variable selection and preprocessing method. The first set preprocesses traditionally and selects input variables under assumption of traditional preprocessing (NTT and CTT in table3). The second set preprocesses traditionally and selects input variables under assumption of categorical preprocessing (NCT and CCT in table3). The third set preprocesses categorically and selects input variables under assumption of categorical preprocessing (NCC and CCC in table3). With above procedure, we can obtain the net effect of each preprocessing method. Table 3 is a matrix for experiment.

	Name of Model	Variable Selection	Preprocessing Method
GA ANN	NTT	Traditional	Traditional
	NCT	Categorical	Traditional
	NCC	Categorical	Categorical
GA CBR	CTT	Traditional	Traditional
	CCT	Categorical	Traditional
	CCC	Categorical	Categorical

<Table 3> Matrix for Phase I Experiment

Phase II: Comparison of integrated model

In Phase II, we integrate Model NCC and CCC, the model of the best performance in each AI Technique. Among the variations of loosely-coupled models as Medsker classified already, post-processing model is employed, regarding an integrated model [8]. In this type of architecture, the one system can perform data manipulation and classify inputs. The other system then performs forecasting and error smoothing. First of integrated model is "Out-of-the-Model Integration" or "GAANN-GACBR". It combines each output of GAANN and GACBR. The combination method is to use the arithmetic mean of outputs of each stand-alone model as an output of integrated model. The second is "In-Model Integration" or "GACBRNN" and "GANNCBR". GACBRNN Model is to add the output of GAANN to GACBR as additional variable and then implement CBR process. In addition, GANNCBR Model is to add the output of GACBR to GAANN as additional input variable

and then implement ANN process.

Results

In Phase I, the followings are the average hit ratio of each model and the preprocessing method. BM in table 4 is the benchmark model, which assumes that the pattern of next day takes the same pattern of current day.

	Name of Model	Training & Testing	Hold-out
Benchmark	BM	N/A	52.67
GAANN	NTT	62.50	57.33
	NCT	56.67	48.00
	NCC	76.33	73.33
GACBR	CTT	N/A	67.33
	CCT	N/A	58.67
	CCC	N/A	76.67

<Table 4> Average Hit Ratio

(T: Traditional Preprocessing, C: Categorical Preprocessing)

We examined the statistical significance test by Two-Sample Test for Proportions. Table 5 is the results.

	NTT	NCT	NCC	CTT	CCT	CCC
BM	0.812	0.808	3.707 **	2.593 **	1.046	4.348 **
NTT		1.619	2.912 **	.787	0234	3.561 **
NCT			4.491 **	3.389 **	.852	5.124 **
NCC				1.138	2.681 **	0.667
CTT					1.555	1.800 *
CCT						3.333 **

<Table 5> Results of Statistical Significance Test (Z-value)

(*: Significant at the 10% level / ** : Significant at the 1% level)

In phase II, the followings are average hit ratio of each model.

Model	Average Hit ratio
GAANN(Stand alone Model)	73.33
GACBR(Stand alone Model)	76.67
GAANN-GACBR(Integrated Model)	75.00
GACBRNN(Integrated Model)	76.00
GANNCBR(Integrated Model)	73.33

<Table 6> Final Average Hit Ratio

The research outcomes may be summarized as follows;

First, genetic algorithms are a useful and effective method to select input variables for AI techniques. Second, the results of the experiment with categorical data preprocessing significantly outperforms that with traditional data preprocessing in forecasting up/down fluctuant direction of index futures price. Third, GACBR outperforms GAANN, Forth, the integration of GA, CBR and ANN (GAANN-GACBR, GACBRNN and GANNCBR) provide worse results than GACBR.

Future Research Issues

First, advanced integration method of models is needed. To justify the integrated model, each technique gives mutual complement to it.

Second, to develop a model producing segmented results (for example, fluctuation rate etc.) is needed. It can give effectiveness to information users.

Third, it is desirable to extract reasonable trading rules and model Decision Support System using these rules.

Reference

[1] Buta, P., "Mining for Financial Knowledge with CBR", AI Expert, February, 1994.

[2] Goldberg, D.E., "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison-Wesley Publishing Company, Inc., 1989.

[3] Kim, Kyoung-jae, "Prediction of Stock Index Futures Price using AI Techniques with Categorical Preprocessing: Case of KOSPI 200 Futures Market", Masters Thesis in KAIST, 1997.2.

[4] Kim, Steven H., "Knowledge Based Prediction of Nonlinear Phenomena", Work Paper, KAIST, Korea, 1995.

[5] Kim, S.H. & Kang, D.S., "Implicit versus Explicit Learning for Forecasting: Case Study in Intraday Stock Index Prediction", Working Paper, 1996.

[6] Kolb, R.W. & Hamada, R.S., 'Understanding Futures Markets", Scott, Foreman and Company, 1988.

[7] Kolodner, J., "Case-Based Reasoning", Morgan Kaufmann Publishers, Inc., 1993.

[8] Medsker, L.R., "Hybrid Neural Network and Expert Systems", Kluwer Academic Publishers, 1994.

[9] Van Eyden, R.J., "The Evaluation of Certain Technical Analysis Techniques to Determine the Future Market Price of Shares on the Johannesburg Stock Exchange", Unpublished M.Com Thesis, University of Pretoria, 1991.

[10] Van Eyden, R.J., "The Application of Neural Networks in the Forecasting of Share Prices", Finance & Technology, 1996.

[11] Weiss, S. & Kulikowski, C. "Computer Systems That Learn", Morgan Kaufmann Publishers, Inc., 1991.