Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm
2015 Vol.5 (S1), pp. 5335-5348/Mahmoudi et al.
Review Article

ESTIMATING STOCK MARKET INDEX VARIATIONS PATTERN USING GENETIC ALGORITHM AND SELECTED ARTIFICIAL INTELLIGENCE METHODS

[*]Monir Mahmoudi[1], Rasoul Nouralsana[2] and Fatemeh Rasouli[3]
[1]Department of Science, College of Industrial Engineering, South Tehran Branch, Islamic Azad University
[2]Department of Industrial and Systems Engineering, College of Industrial Engineering, University of Elmosanat (Tehran-iran)
**Author for Correspondence*

ABSTRACT
Predicting the stock price index and the direction it moves has been the subject of many experimental researches, but much of the outcome has been related to the developed financial markets; regarding the developing markets, researches are not many. Considering the high analytical power of data processing and its incomparable capability, it can be used in the analyses of many problems including forecasting cases. The main objective of this research study is to develop a prediction method that can forecast the descending/ascending direction of the future stock price. To model the data processing techniques, use has been made of genetic algorithm (GA) to find the optimal variables; to solve the model, such artificial intelligence approaches as "artificial neural network" (ANN), "optimized vector machine", "decision tree" (DT) with AUC precision index, area under ROC graph, and F-index have been used. In the present paper, use has been made of the data related to Istanbul Stock Exchange (ISE) index. The period between 1997 and 2007 has been the time when the data were gathered amounting to 2733 records per day. Then, the values related to 10 calculated indexes have been used as the inputs to the genetic algorithm and the forecasting related to the stock movement direction has been modeled. Ultimately, using the optimized vector machine, the best return has been determined.

Keywords: *Data Processing; Artificial Intelligence; Genetic Algorithm; Stock Price Index; Optimized Vector Machine*

INTRODUCTION
Stock price index shows the general economic situation in a country; its rise means economic improvement and its fall shows crisis and recession. Its forecasting, therefore, can be beneficial to investors, industry owners, and even market analysts. The stock market is generally affected by such macro-economic elements as political events, companies' policies, general economic conditions, institutional investors and their psychology, changes in other stock markets, and so on. In the past years, those active in this market have used classic methods to predict stock price, but considering the continued progress in such methods as the artificial intelligence, heuristic and metaheuristic algorithms, genetic algorithm, support vector machine, artificial neural network, and fuzzy artificial neural network, they have found ever-increasing applications as regards the forecasting of the stock price index. In this research, use has been made of the highly efficient data processing technology and its unparalleled processing capability for problem analysis in Turkey stock market. This technology makes use of the artificial intelligence and metaheuristic algorithms to do forecasting. The method of this research has been based on the crisp process, and the modeling includes two main steps of feature selection (by GA) and modeling (by classification methods). The data used are related to the daily stock price change direction (in ISE-100 index) from 1997 to 2007.
In this research, use has been made of 10 variables; in part 1, the variables affecting the objective variable were found using GA. This way, the data set variables were reduced and the model precision was enhanced. In part 2, after determining every model's parameters, different classification methods such as neural networks, support vector machine, decision tree, etc were used on the data set and finally the

results obtained from every model's evaluation were explained. To develop the model, use has been made of 2733 data records of Istanbul Stock Exchange index. Modeling through ANN, SVM, and BPN resulted in precisions of respectively 75.74%, between 59.35% and 71.43%, and 75.74%; therefore, the ANN model with a higher precision was selected.

Yakup *et al.,* (2011) proposed their model using such methods as the NN, SVM, gradient with the AUC measuring precision and classification with one time validation. To predict the ISE index movement direction, Diler (2003) has used neural networks; the indexes used have been RSI and (MACD) MA momentum. The results of these researches show that the ISE price index movement direction is predictable with a precision of 60.8%. Chen *et al.,* (2003) tried to predict the movement direction of Taiwan stock index; they used the possibilistic neural networks for their purpose. The statistical performance of PNN forecasting was compared with such extended momentum methods as Kalman filter and stochastic step, and it was shown that PNN had a more powerful forecasting capability (Chen, 2003). Avci *et al.,* (2007) compared the ANN models' efficiency with that of SVM in predicting ISE-100 National index and showed that the former worked much better (75.74%) than SVM (71.52%). Enke *et al.,* worked on predicting stock exchange through predicting one stock or index that focuses on predicting the level (value) of the prices of the future market and, also, the exchange price movement direction; they introduced a 3-step forecasting system consisting of: 1) a multi-agent regression analysis, 2) a type 2 fuzzy clustering, and 3) a type 2 fuzzy NN to predict the stock exchange. The model they obtained performs better than the traditional ones in predicting the stock price (Enke, 2011). Khansa *et al.,* made use of auto regression analyses and ANN as complementary methods to study the relation between the stock return of information security companies and the severity of malicious attacks and computed the severity level of a number of such attacks. A major part of this work is the time-delayed ANN model that facilitates the forecasting of stock return and is, especially, beneficial, as an investment decision making support system for protection funds and other investors whose bonds are subject to losing their market value due to malicious attacks (Khansa, 2011). Sung *et al.,* worked to analyze the dependence law to predict the variations in (South) Korea Composite Stock Price Index (KOPSI) based on the time series data of different indexes of the global stock markets.

Results have shown that KOPSI tends to continue the same direction as Europe and US stock market indexes (Sung, 2011). Gonzales et al. used, in their research, feed-forward, 3-layer, ANNs and RSI technical index to improve business systems by introducing the CAST technique (Gonzales, 2011). Huang et al. tested an improved NN model on IXIC, DJI, HIS, SBI, SAI, and SP 500 data, and presented a data processing technique based on the promising learning machine to find the forecasting relations among many financial and economic variables (Huang, 2011). Liao *et al.,* compared capital assets price model (CAPM) with the French 3-agent Fama model to check the forecasting power of single and multi variable NN models in predicting Shanghai stock price movement direction; results showed that ANNs worked better (Liao, 2011). Pincak compared the PMBSI model performance with SVM and ANN in artificial and financial-time series. The first model is based on the correlation function as a constant and the second is a program based on the deviation from PMBCS. The first predicts the annual return, but cannot predict the foreign exchange market behavior with good return (Pincak, 2013). Kazem *et al.,* proposed, in their paper, a forecasting model based on turbulent surveying, glowworm algorithm, and support vector regression (SVR) to predict the stock market price. The model has 3 steps of coordination arrangements, irregular glowworm algorithm, and optimized SVR. It presents its best performance based on a dual error criterion (mean squared error- MSE – and mean absolute percent error- MAPE) (Kazem, 2013). Tiknor presented a regular Bayesian method of ANNs as a new approach for predicting financial markets behavior. To find how effective the model is, he performed some tests on Microsoft and Goldman Sox stocks and showed that the proposed model, a very advanced one, worked with no needs to data processing, seasonal tests, and analyses (Tiknor, 2013). Park *et al.,* proposed a stock forecasting method using a semi-supervision learning (SSL) algorithm to overcome the constraints (Park, 2013). Yuan presented a new learning model called polynomial smooth support vector machine (PSSVM). The optimum forecasting parameters were found after the problem was solved using Broyden-Fletcher-

Goldfarb-Shanno (BFGS) method. Results showed that the proposed learning model was powerful and effective (Yuan, 2013).

Model Structure

The model proposed in this research has been taken from the original crisp model and is shown in Figure 1. Modeling consists of 2 main steps of feature selection through GA and modeling through the use of some classification models. All the steps are explained in detail in Figure 1 and the results found from the evaluation of every model are separately studied at the end of the paper through some tables and figures.

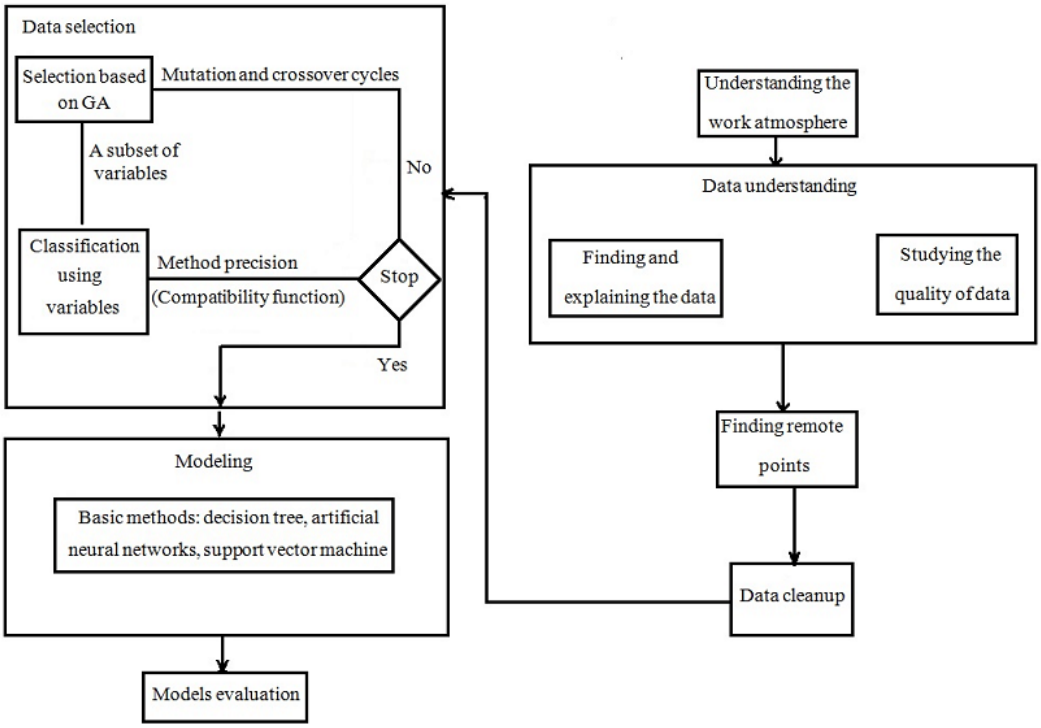

Figure 1: Research execution model

Data Understanding

The data used in this paper are related to the daily price variation movement direction of ISE-100 National Index from 1977 to 2007 amounting to 2733 records (1440 ascending and 1293 descending – 52.7 and 47.3% respectively). These historical data have been found through Matrixgold 2.4.0 Technical Analysis Module, produced by Matrix Information Presenting Company. The number of samples per year in the data set is given in Table 1.

Table 1: Number of records in different years

Year	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	Total No.
Ascending (No.)	126	131	147	142	134	123	123	111	133	124	146	1440
Ascending (%)	50/2	52/4	57/9	57	54/5	48/8	49/6	44/9	56/4	50	57/9	52/7
Descending (No.)	125	119	107	107	112	129	125	136	103	124	106	1293
Descending (%)	49/8	47/6	42/1	43	45/5	51/2	50/4	55/1	43/6	50	42/1	47/3
Total	251	250	254	249	246	252	248	247	236	248	252	2733

Table 2 shows the name, type, and a brief explanation of variables in the data set

Table 2: Technical indexes' applied formulae

Index name	Formula
Moving average, 10-day simple	$\frac{C_t + C_{t-1} + \cdots + C_{t-10}}{10}$
Moving average, 10-day weighted	$\frac{((n) \times C_t + (n-1) \times C_{t-1} + \cdots + C_{10}}{n + (n-1) + \cdots + 1}$
Amount of movement	$C_t - C_{t-n}$
K% stochastic	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$
D% stochastic	$\frac{\sum_{i=0}^{n-1} K_{t-1}\%}{n}$
RSI (Relative Strength Index)	$100 - \frac{100}{(1 + \sum_{i=0}^{n-1} Up_{t-1/n}) / \sum_{i=0}^{n-1} DW_{t-1}/n}$
Moving Average Convergence Divergence MACD	$MACD\ (n)_{t-1} + 2/n + 1 \times (DIFF_t - MACD\ (n)_{t-1})$
Larry Williams R%	$\frac{H_n - C_t}{H_n - L_n} \times 100$
A/D oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t} \times 100$
CCI (Commodity Channel Index)	$\frac{M_t - SM_t}{0.015D_t} \times 100$

Data Explanation

Variables' features have been explained using Medler software and some simple and double variable statistics. Table 3 shows the results obtained from single variable analysis using some such indexes as average, standard deviation, and so on.

Table 3: Data statistical explanation

Variable name	Variable type	Min.	Max.	Range	Average	Standard deviation	Middle	Mode
Simple MA	Continuous	951.1	57155.83	56204.73	17867.861	14717.273	12145.355	1606.7
Weighted MA	Continuous	961.87	57450.36	56488.49	17897.526	14739.612	12166.71	961.87
Momentum	Continuous	61.51	159.3	97.79	101.949	9.65	101.595	103.33
S (K %)	Continuous	6.86	99.34	92.48	56.813	24.733	59.405	64.83
S (D %)	Continuous	11.81	97.3	85.49	56.821	19.349	57.095	53.93
RSI	Continuous	14.4	96.25	81.85	54.493	13.1	53.99	42.54
MACD (26.12)	Continuous	-2117.35	2075.33	4192.68	138.095	508.277	81.56	-348.54
Williams	Continuous	-100	0	100	-41.728	30.259	-38.945	0
A/D	Continuous	108446900	80498970000	80390523100	21065003370	23491507886	10297390000	149735500
CCI	Continuous	-323.22	288.21	611.43	12.933	86.996	19.385	1.88

For example, since movement size variable oscillates around 100 (it is 101.949), it can be claimed that the rise or fall of the prices in this data set has been almost the same for the records.

Double-variable Explanation of the Data

Pierson Test, well known to check variables' dependence/independence with respect to one another, was performed on the variables; correlation rates are given in Table 4.

Table 4: Results of Pierson Test performed on the variables of the present research

Variable 1	Variable 2	Correlation
Simple MA	A/D	0/965
MOMENTUM	S(%D)	0/754
MOMENTUM	CCI	0/772
S(%D)	RSI	0/777
S(%D)	A/D	0/071

As shown, some variables are highly correlated. It is to be noted that in some data processing methods, the default is that independent variables are not closely correlated. According to Table 4, it is obvious that S (D %) is closely related to RSI, but less to A/D index.

Data Description

Some graphs can show the features of and relations among the variables. Figure 2, e.g., shows Williams index for 2 groups of ascending and descending records.

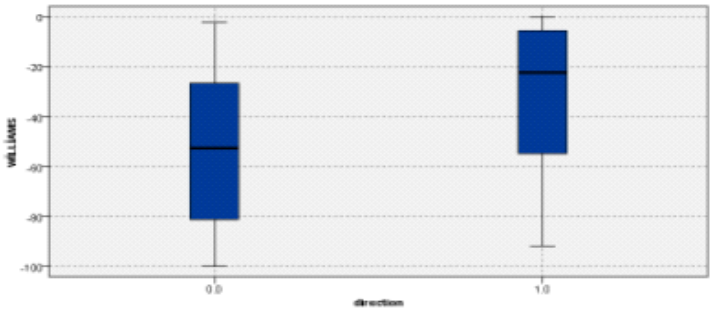

Figure 2: Box graph

The above graph shows that the ascending records have been more inclined towards higher values of Williams index the middle of which (for variables) is close to 20. On the other hand, the index middle for descending ones is close to -50 which means that they have been inclined more towards lower values of Williams' index. Fig. 3 shows the records' dispersion considering CCI and weighted MA indexes. The number of records in every part of Figure 3 has been specified with dark blue.

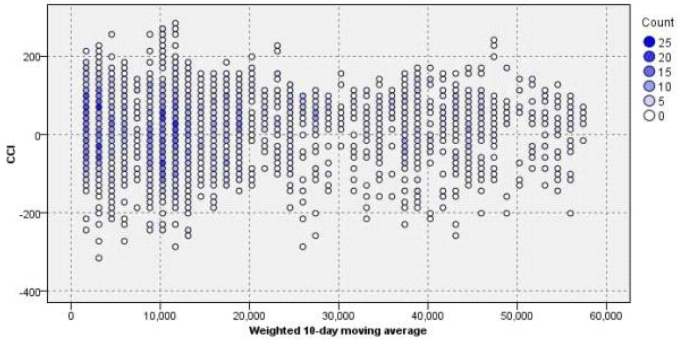

Figure 3: Records' dispersion

As shown, most records have had smaller weighted MA because the samples on the left side of the figure are darker. The figure also shows that samples having averages around 30000 have less CCI dispersion. It can be generally stated that most of the available records have been dispersed around CCI = 0. Fig. 4 shows the dispersion related to two variables of stochastic K % and simple MA; the ascending samples are in red and the descending ones are in blue.

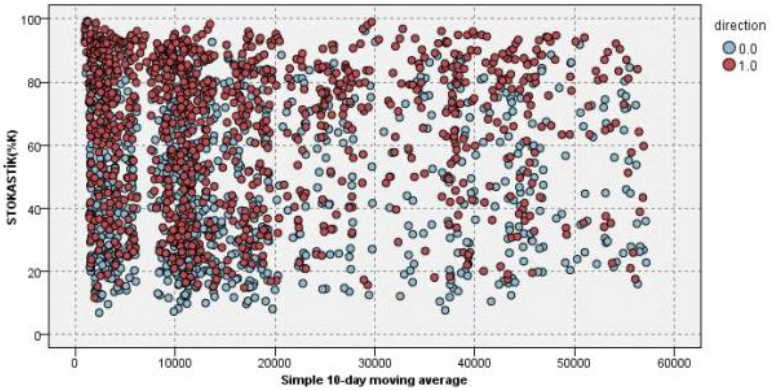

Figure 4: Dispersion considering the objective variable

As shown, most of the samples have simple MAs smaller than 20000 units

Data Preparation

Before entering the modeling phase, the data should be cleaned up first. To find the remote points, use was made of Shewharts' control region graph the equation of which is as follows:

معادله در متن فارسی داده نشده است

where …. and …. are respectively the average and standard deviation of the variable in question. This control region has been used to find the remote points in all the variables; in this region, ….=3 is to find the number of remote points and ….=6 is to find highly remote points. Table 5 shows the results of this method used on the data.

Table 5: Remote and highly remote points found through Shewharts' control region graph.

Variable name	Remote points	Highly remote points
Simple 10- day moving average	0	0
Weighted 10- day moving average	0	0
MOMENTUM	33	1
STOKASTİK(%K)	0	0
STOKASTİK (%D)	0	0
RSI	4	0
MACD(26.12)	61	0
WİLLİAMS	0	0
A/D	0	0
CCI	12	0

Table 5 shows that MACD variable contains the highest number of remote points (61 points). All the remote and highly remote points have been omitted in Table 5 and replaced with every variable's middle.

Modeling

This section consists of 2 main parts: 1) finding, through GA, variables that affect the objective variable and 2) applying different classification methods such as NN, SVM, DT, etc., to the research data set after determining the parameters of every model.

Feature Selection

In this step, use has been made of such advanced search methods as GA to find the best set of variables. In selecting a subset of variables through GA, mutation means selection or rejection of variables in every execution. Crossover, too, means changing the used variables. In this algorithm, selections are done through different methods (e.g. rotating wheel).

Feature selection based on the GA compatibility function, which shows optimality, is a classifying model. This means that when the primary set of variables has been selected based on the adjusted probabilities, modeling is done through the use of this set and a classification method and then the model precision is evaluated. Next, selection, mutation, and crossover will continue until an optimum is reached in the constrained model. Fig. 5 shows the operation of this method.

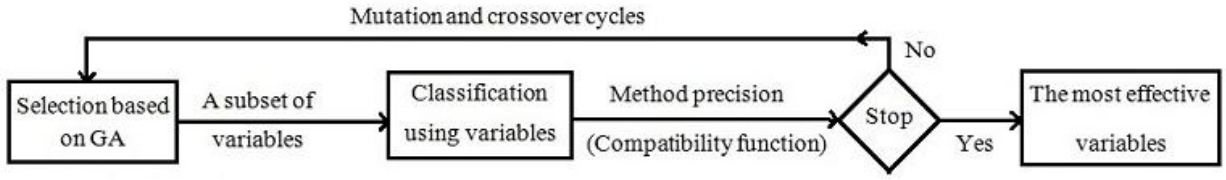

Figure 5: Feature selection through GA

In this paper, to check the selection correctness, use has been made of DT (with the information return index) for classification. In this method, the society has 20 members, mutation probability is 0.033, crossover probability is 0.6, safety index (for pruning) is 0.25, and average number of samples in every leaf is 2. Out of 10 variables, 7 were selected in this step (Table 6) which will be used later in the modeling step.

Table 6: Results of using GA in finding effective variables

GA selection	Index
Not selected	10-day simple MA
Not selected	10-day weighted MA
Not selected	Movement size
Selected	Stochastic K %
Selected	Stochastic D %
Selected	RSI
Selected	MACD
Selected	Williams R %
Selected	A/D
Selected	CCI

Model Parameters Adjustments

Using non-replacing stochastic sampling, 10 % of the data were used to adjust the parameters; to find the optimum ones use was made of CV parameter function in Veca software. This function is capable of testing different combinations of parameters and specifying the best one considering the model precision. Parameters adjusted for different models (with the optimum ones) are given in the following parts. To validate the modeling, use was made of 10 cross-validation procedures wherein 90 % of the data were first used to train the model and the remaining 10 % to test the model, and then 10 % of the 90 % training data were separated and used for testing while the 10 % used in the first step were used for the purpose of training in the second round. This was repeated 10 times so as to make use of all the data records in both training as well testing the model.

Nearest Neighborhood

This method has only one parameter (K) which shows the number of neighboring points for modeling. In this research K was between 1 and 20. Results: model precision: 69.6 %, K: 11.

Decision Tree (J48)

The adjustable parameters for this model are: 1) C which shows the safety factor for pruning, and 2) M which indicates the minimum number of samples in every leaf to stop the tree from growing; M ranges from 10 to 100 (with 10 steps) and C ranges from 0.1 to 0.4 (with 7 steps). Results: M: 40, C: 0.25, precision: 63.73 %.

Artificial Neural Network

The 3 parameters that affect the output precision of these networks include momentum (M), number of executions (N), and learning (L). In this research, M ranges from 0.2 to 0.5 (with 4 steps), N ranges from 400 to 600 (with 3 steps), and L ranges from 0.2 to 0.4 (with 5 steps). Results: M, L: 0.4, N: 400, precision: 74.35 %.

Optimizes Support Vector Machine

In this method, there are 3 core functions including POLY, PUK, and RBF. The only adjustable parameter in this model is C (model complexity index) which was adjusted between 0 and 30 (with 7 steps) for all the 3 core functions. Results: (for POLY and PUK): C: 5, precision: 74.72 % and 71.06 %, respectively, and (for RBF) C: 25, precision: 71.79 %.

Final Modeling

Considering the parameters found in the previous steps (clean-up and variable reduction), 7 models were applied in the "Experimenter" part of Veka software. In this step, use was made of 10 data set validations and 10 times execution of every model; every model was run 100 times.

Since in every modeling data enter the model randomly, the result by different methods in every execution are different too; the smaller the differences or the deviations, the more precise the model. Table 7 shows the results for all 7 models obtained as regards the precision index. To simplify the efficiency comparison in Fig. 6, model precision has been shown for the standard deviation in every execution.

Table 7: Evaluation with precision index

Model	ACC	
	Avg.	Std.
NB	62.27	2.7
SMO – PLN	77.97	2.24
SMO – PUK	80.05	2.3
SMO – RBF	77.09	2.26
MLP	79.21	2.7
J 48	78.43	2.39
KNN	75.58	2.64

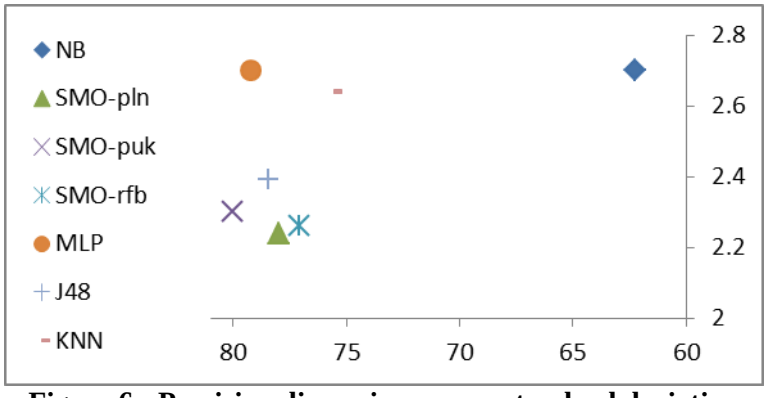

Figure 6: Precision dispersion versus standard deviation

The optimized SVM with PUK core function and a precision of 80.05 % has been more precise than other models; therefore, SMO – PUK method is lying in the lower part of the right side in Figure 6. On the other hand, the NN methods too have rather high precisions, but as shown in Figure 6, its standard deviation in 100 executions is more than those of the other methods which is considered as the method's weak point. The simple Bayesian and nearest neighborhood methods have shown to be weak as regards the precision index; other methods based on optimized SVM have averagely had appropriate, close to each other precision indices. Fig. 7 shows the area under ROC (AUC) graph index considering the deviations found for it in all models. This index shows the area under a graph drawn using the ratio of the correct positive to wrong positive indices.

Table 8: Evaluation with ROC index

Model	**ROC**	
	Avg.	Std.
NB	0.69	0.03
SMO – PLN	0.78	0.02
SMO – PUK	0.8	0.02
SMO – RBF	0.77	0.02
MLP	0.88	0.02
J 48	0.84	0.02
KNN	0.84	0.03

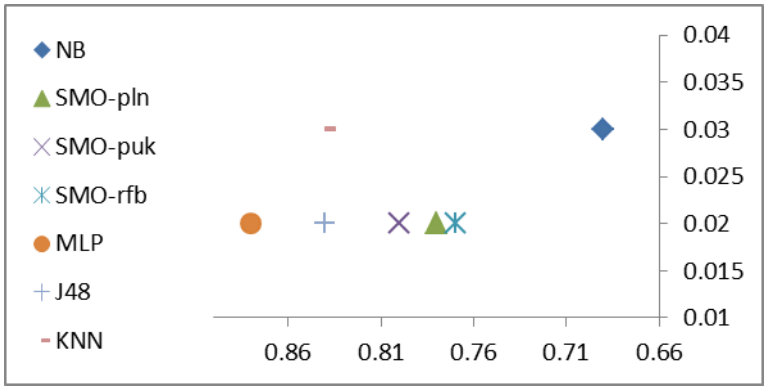

Figure 7: ROC dispersion versus standard deviation

In the NN method, the ROC index was found to be 0.88 which is more optimal than those found in other methods. But, the optimized SVM with PUK core function stands 4th among other models as regards the ROC index. Since K-nearest neighborhood method has had an inappropriate performance regarding the precision index, it has shown a better result in ROC index compared to other models; DT too has had a ROC index of 0.84 which is the second best after the NN method.

Table 9: Evaluation with F-Measure

Model	**F-Measure**	
	Avg.	Std.
NB	0.64	0.03
SMO – PLN	0.79	0.02
SMO – PUK	0.81	0.02
SMO – RBF	0.79	0.02
MLP	0.8	0.03
J 48	0.79	0.02
KNN	0.76	0.03

It is worth mentioning that the simple Bayesian method has not had an appropriate performance as regards the ROC index either. The F- Measure has been used in this research because it is the result of the combination of the precision index (ratio of the correct predicted positives to the whole predicted positives – the majority class) and the recall index (ratio of the correct predicted positives to the whole real positives). Table 9 presents the results of this index for different models.

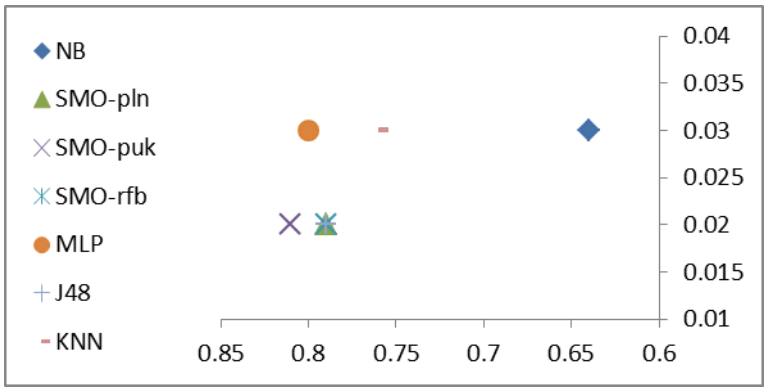

Figure 8: F-Measure dispersion versus standard deviation

F-Measure results show that the optimized SVM with PUK core function and an index of 0.81 has had a better performance than the other methods. The ANN method, on the other hand, has gained 0.8 for this index, but its standard deviation after 100 executions of the model has been more than that of the above method. In this index too, it is clear that the simple Bayesian and K-nearest neighborhood have had the worst performance as regards this index (similar to the precision index).

Analysis Based on T-Test

In this research, the models were made for 100 times, and, therefore, we can find 100 different values for every evaluating index. Here, use has been made of the T-test to check the meaningfulness of the difference between the indices values found for every model. In the following Tables, the "Failure" column, considering the T-test, shows the number of models wherein the value of the index in question is meaningfully worse than the model in question. The "Success" column too shows the number of models compared to which the model in question has a meaningful difference. The "Difference" column too shows the difference between the "Success" and "Failure" columns. It is worth mentioning here that all the models in Table 10 have been arranged based on the "Difference" column for the 3 studied indices in the descending order, and the models that have higher places in this Table have had better performances, in the index in question, than other models regarding the T-test. Table 10 shows the T-test results based on the existing decision index.

Table 10: T-test results for the existing decision index

Model	Failure	Success	Difference
SMO-puk	0	4	4
MLP	0	3	3
J48	0	2	2
SMO-pln	1	3	2
SMO-rfb	3	1	-2
KNN	4	1	-3
NB	6	0	-6

As shown, the optimized SVM model with PUK core function has a meaningfully more precision index compared to the other 4 models used in this research. The ANN model stands second and the simple Bayesian and K-nearest neighborhood too have performed meaningfully weaker than respectively the

other 6 and 3 models (it is to be noted that Table 10 does not show whether the model in question has performed better or worse than other models). Table 11 and Table 12 show ROC index and F-Measure T-tests respectively.

Table 11: ROC index T-test

Model	Failure	Success	Difference
MLP	0	6	6
KNN	1	4	3
J48	1	4	3
SMO-puk	3	3	0
SMO-pln	4	2	-2
SMO-rfb	5	1	-4
NB	6	0	-6

Table 12: F-Measure T-test

Model	Failure	Success	Difference
SMO-puk	0	3	3
J48	0	2	2
MLP	0	2	2
SMO-pln	0	2	2
SMO-rfb	1	2	1
KNN	5	1	-4
NB	6	0	-6

Table 11 shows that the ANN model has performed meaningfully better than all the other 6 models. The K-nearest neighborhood and DT models stand second; they have performed better than other 4 models as regards this index (they have performed weaker than only one model). Table 12 shows the T-test results for F-Measure. As shown, the optimized SVM with PUK core function has performed meaningfully better than the other 3 models regarding the F-Measure. DT, ANN, and optimized SVM with POLY core function stand next; they have performed better than the other 2 models in this research.

RESULTS AND DISCUSSION

Results

In this paper, the research procedure was explained based on the crisp process and the case study was carried out using real data of Istanbul Stock Exchange (ISE). Explanation and illustration of the indices in the data set were carried out after they were introduced. Some of the results obtained from the statistical explanation and illustration of the data can be summarized as follows:

The average value of – 41.728 for R % (William %) means that, on average, the samples, considering this index, have lain neither in the sale saturation region (80-100 %) nor in the purchase region (0-20 %). Since the values for A/D index are close to the minimum, it can be concluded that the higher values of this variable have been probably more inclined to the descending direction and sale. The S (D %) index highly correlates with the RSI index, but its correlation is weak with the A/D index. The samples the weighted moving averages of which are around 30000 have less CCI dispersion.

After data preparation and omission of the remote values based on Shewhart range control, the data entered the modeling phase. First, 7 out of 10 independent variables were selected as the effective ones using GA (with DT fitness function), and entered the modeling phase; then, the data set in question entered the 4 classification models. First, the optimal parameters were selected for every model. Some results obtained from parameter adjustment for every model are mentioned next.

K-nearest neighborhood: The more are the Ks with respect to the default, the more will be the model precision.

Decision tree (DT): The more are the samples in a leaf (to stop the tree from growing) with respect to the total records, the more will be the precision.
Artificial neural networks (ANNs): It was found in this research that the rate of learning should be more than the default value so that the network may reach the optimal solution sooner.
Optimized support vector machine (OSVM): Considering the type of the core function, parameter C can adopt different values to enhance precision.
Next, using every model's optimum parameter, modeling was done through 10 times cross-validations and 10 executions. Results of studying the average precision and its standard deviation in 100 times modeling for all the models are briefly given in the following lines. The optimized SVM with PUK core function and a precision of 80.05 % has had a better performance than all the other models. The NN model too has gained a high precision, but a major weak point of this model is its high deviation found in 100 times execution. The simple Bayesian model, among all the methods used has been the weakest classification model. The models' investigation results, as regards the AUC index, are as follows:
The ANN model has had better performance than other models. K-nearest neighborhood has not had a good performance regarding the precision index, but has performed better in this index compared with other models The optimized SVM with PUK core function stands 4th among all the models as regards the ROC index. Models' investigation results regarding F-Measure are as follows:
The optimized SVM with PUK core function and an F-Measure of 0.81 has performed better than the other models. The simple Bayesian and K-nearest neighborhood have had the weakest performance in this index (as was the case with the precision index).
Next, use has been made of T-test to investigate the meaningfulness of the difference among every model's index value. Results have shown that the optimized SVM with PUK core function is the best as regards precision and F-Measure. The ANN too has been the best as regards the area under the AUC graph. Therefore, if the ratio of the correct to incorrect positive values is important, it is better to use the ANN model, or else, making use of the optimized SVM with PUK core would be the best. Now, to clarify the results and findings of the present research, they have been compared with those of other researches. Table 13 gives a comparison of the modeling precision found in the proposed paper and in some other related researches.

Table 13: Comparison of the results of the present study with those of some other related researches

Author	Year published	Model	Precision
Diler	2003	NN	%60/81
Altay	2005	NN	%57/80
Kara	2011	NN	%75/74
Proposed model	-	SVM+GA	%80/05

Suggestions for Future Studies
It is worth mentioning that modeling validation has not been done properly in other papers. As an example, the data set used in the present research was taken from a paper by Kara *et al.,* (2011) who had done validation only once. But, in this research, use has been made of 10 times cross validations resulting in an enhanced precision. Table 13 shows that the proposed model is more precise than other models and has been able to achieve the main goal of the research. Since the results obtained through modeling of the present study indicate precision improvement in predicting the direction of the stock exchange index movement, the following are suggestions made for the development of the results of this research by other researchers:
- Using collective classification methods – knowing that not many researches have made use of such collective classification methods as Begging and Boosting, it seems necessary that future modeling efforts make use of these methods to enhance the modeling precision.

- Using other feature selection methods – it is necessary that in future, researchers make use of such other optimization methods as Ants Colony and Particle Swarm to do more effective feature selection.
- Using domestic data sets – it is necessary that domestic data sets be used to check the proposed model's efficiency and capability.

REFERENCES

Abu-Mostafa YS and Atiya AF (1996). Introduction to financial forecasting. *Applied Intelligence* **6**(3) 205–213.

Ahmad Kazem, Ebrahim Sharifi, Farookh Khadeer Hussain, Morteza Saberi and Omar Khadeer Hussain (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing* **13** 947–958.

Atsalakis GS and Valavanis KP (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications* **36**(3) 5932–5941.

Avci E (2007). Forecasting daily and sessional returns of the ISE-100 index with neural network models. *Journal of Dogus University* **8**(2) 128–142.

Chu HH, Chen TL, Cheng CH and Huang CC (2009). Fuzzy dual-factor time-series for stock index forecasting. *Expert Systems with Applications* **36**(1) 165–171.

David Enke, Manfred Grauer and Nijat Mehdiyev (2011). Stock Market Prediction with Multiple Regression,Fuzzy Type-2 Clustering and Neural Networks. *Procedia Computer Science* **6** 201–206.

Diler AI (2003). Predicting direction of ISE national-100 index with back propagation trained neural network. *Journal of Istanbul Stock Exchange* **7**(25–26) 65–81.

Fagner A de Oliveira, Cristiane N Nobre and Luis E Zárate (2013). Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index – Case study of PETR4, Petrobras, Brazil. *Expert Systems with Applications* **40** 7596–7606.

Hsu SH, Hsieh JJPA, Chih TC and Hsu KC (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications* **36**(4) 7947–7951.

Huang CL and Tsai CY (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications* **36**(2) 1529–1539.

Huang W, Nakamori Y and Wang SY (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research* **32** 2513–2522.

Jonathan L Ticknor (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications* **40** 5501–5506.

Kanghee Park and Hyunjung Shin (2013). Stock price prediction based on a complex interrelation network of economic factors. *Engineering Applications of Artificial Intelligence* **26** 1550–1561.

Kara Y *et al*., (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications* **38** 5311–5319.

Lara Khansa and Divakaran Liginlal (2011). Redicting stock market returns from malicious attacks: A comparative analysis of vector autoregression and time-delayed neural networks. *Decision Support Systems* **51** 745–759.

Liao Zhe and Wang Jun (2011). Forecasting model of global stock index by stochastic time effective neural network. *Expert Systems with Applications* **37** 14026-14036.

Md. Rafiul Hassan, Kotagiri Ramamohanarao, Joarder Kamruzzaman, Mustafizur Rahman and Maruf Hossain M (2013). A HMM-based adaptive fuzzy inference system for stock market forecasting. *Neurocomputing* **104** 10–25.

Pincak R (2013). The string prediction models as invariants of time series in the forex market. *Physica A* **392** 6414–6426.

Sung Hoon Na and So Young Sohn (2011). Forecasting changes in Korea Composite Stock Price Index (KOSPI) using association rules. *Expert Systems with Applications* **38** 9046–9049.

Wensheng Dai, Jui-Yu Wu and Chi-Jie Lu (2012). Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes. *Expert Systems with Applications* **39** 4444–4452.
Xu X, Zhou C and Wang Z (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications* **36** 2625–2632.
Yakup Kara, Melek Acar Boyacioglu and Ömer Kaan Baykan (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications* **38** 5311–5319.
Yubo Yuan (2013). Forecasting the movement direction of exchange rate with polynomial smooth support vector machine. *Mathematical and Computer Modeling* **57** 932–944.