Proc. of the Intl. Conf. on Advances in Human Science, Economics and Social Study-- AHES 2014

ISBN: 978-981-07-8859-9

Thai Stock Index directional prediction using higher order differences and higher order lag inputs

Phattradanai Samurwong, Krung Sinapiromsaran

Abstract— **Stock directional classification technique proposed in this paper is to determine the return direction of stock market. Some previous works used a support vector machine to predict the return direction of the stock market index based on first order difference, stock return, as inputs. Our paper will use higher order differences and higher order lags as inputs for a support vector machine. By using higher order differences and higher order lags of time series data, the experimental results show the improvement of the prediction accuracy with respect to the number of lags and the number of difference orders. In addition, results of this technique show better daily directional prediction accuracy comparing with the other models.**

Keywords*— *Stock index prediction, support vector machine, order difference, order lag.

I. Introduction

Stock prediction is one of the most interesting topics in both data mining field and in finance field. There are many algorithms and approaches that are used in the daily stock prediction. For examples, Fernandez-Rodriguez et al. (2000)[4] used neural network model to predict Madrid stock exchange general index, Harvey et al. (2000)[7] and Halliday (2004)[6] used their neural network models to predict New York stock exchange index, Lendasse et al. (2000)[12] used radial basis function neural network to predict Belgium stock exchange 20 index, Francis E.H. Tay, Lijuan Cao (2001)[18] used support vector machine to predict S&P index, Kim K-j.[11] used support vector machine to predict Korea stock exchange index, Doesken et al. (2005)[3] used a Mamdani fuzzy system and a Takagi-Sugeno fuzzy system to predict New York stock exchange index, S.-H. Hsu et al.(2009)[8] used the Support vector machine(SVM) + Self Organizing Map(SOM) model to predict Hangseng index. Support vector machine is one of the prediction models that is popularly used due to its accuracy in the prediction of stock [8,9].

Support vector machine can be used for distinguishing different characteristics of the data provided by the hyperplane which acts as a linear separator. The best hyperplane is obtained from solving a convex optimization problem [17] to achieve optimal weight and the offset of hyperplane. After the hyperplane is obtained, it can be applied to the unknown dataset from the same specific population. In case of using SVM for the stock return directional prediction, SVM will be able to distinguish between the positive and the negative return (direction) of the stock index from the multiple input variables provided. For instances, Kyoung-jae Kim (2003) [9] predicted the daily positive and negative direction of the Korea Composite Stock Price Index (KOSPI) and Wei Huanga, Yoshiteru Nakamoria, Shou-Yang Wang (2005) [11] predicted the weekly direction (positive and negative) of NIKKEI index using S&P500 index and exchange rates of Japan Yen to US dollar.

Normally, the input variables of the support vector machine, in the prediction of stock, are time series of stock return which are 1st order difference, technical indicators, and fundamental factors. For instances, Stock exchange of Thailand index (SET) by C. Khumyoo (2000) [10] is predicted by using time series of Dowjones, NIKKEI, Hangseng index, etc. Technical indicators are used by Kyoung-jae Kim (2003) in KOSPI directional prediction via SVM [11]. Huanga, Yoshiteru Nakamoria, Shou-Yang Wang (2005) [9] stated that some fundamental and economic factors including interest rates, consumer price index, gross domestic product, etc.; are used in many studies for the financial prediction. In our paper, we are focusing on using higher order differences and higher order lags of the time series data as inputs for a support vector machine.

In econometric time series, the higher order differences are rarely used to construct financial prediction models [13]. In this paper, n^{th} order differences prone to provide more information that can be used in the data mining scheme. Also, a number of lags, in econometric time series, is not likely to be large. Hence, a larger number of lags can be applied to this model in order to extract features that can be used in the directional prediction in SVM.

In this paper, we track order differences of stock indices return using as input variables from the first order difference to the highest order difference that does not decrease the accuracy of SVM. In addition, we apply this technique to generate input from commodities price return and USD/THB exchange rates return. Also, lag of stock indices return and the related factors, are also considered.

Phattradanai Samurwong*(Author)*
Faculty of Commerce and Accountancy, Chulalongkorn University
Thailand

Krung Sinapiromsaran *(Co-Author)*
Faculty of Science, Chulalongkorn University
Thailand

There are two major indices in the Thai stock market that are mostly referred. The first one is the Stock Exchange of Thailand index, SET index, which represents all Thai stocks. The second one is SET50 which represents 50 largest market capital stocks. In this paper, SET index return directions will be used as the target output in the directional prediction.

II. Methodology

A. *Software used*

According to F. Shafait, M. Reif, C. Kofler, T.M. Breuel (2010) [16], they proposed that, in pattern recognition engineering, Rapidminer has several more advantages than other tools. Those advantages includes the collaboration in user interface which reduce the requirement of expert knowledge, high acceptance in the data mining community, the complete data mining models are constructed and optimized automatically, faster in model training and testing since there is a self-tuning approach, etc. From these advantages, in this paper, we will use Rapidminer [15].

B. *Data Preprocessing*

1) Inputs for the support vector machine

As the input data should be related to the prediction output, which is SET index directional return, thus inputs include time series of stock market indices return, commodities price return, and exchange rates return. These are factors that usually affect SET index [10]. Table I shows the 1st order difference inputs for the directional classification.

TABLE I. THE 1ST ORDER DIFFERENCE INPUTS FOR THE DIRECTIONAL CLASSIFICATION

Input	Formula
$R_{SET_{t-1}}$	$SETclosed_{t-1} - SETclosed_{t-2}$
R_{HSI_t}	$HSIopened_t$- $HSIclosed_{t-1}$
$R_{Dow_{t-1}}$	$DJclosed_{t-1} - DJclosed_{t-2}$
$R_{STI_{t-1}}$	$STIopened_t$- $STIclosed_{t-1}$
$R_{FTSE_{t-1}}$	$FTSEclosed_{t-1} - FTSEclosed_{t-2}$
$R_{Gas_{t-1}}$	$Gasclosed_{t-1} - Gasclosed_{t-2}$
$R_{Gold_{t-1}}$	$Goldclosed_{t-1} - Goldclosed_{t-2}$
$R_{Oil_{t-1}}$	$Oilclosed_{t-1} - Oilclosed_{t-2}$
$R_{THB/USD_{t-1}}$	$THB/USDclosed_{t-1} - THB/USDclosed_{t-2}$

- $R_{SET_{t-1}}$ refers to SET index daily return at time $t-1$ where t refers to today, $t-1$refers to yesterday and $t-2$ refers to the day before yesterday and $SETclosed_{t-1}$ refers to the yesterday closing price of SET index
- R_{HSI_t}, Hangseng index opening return, is calculated from the opening Hangseng Index return of Thailand's time t.
- R_{Dow} refers to Dowjones index daily return $DJclosed$ refers to Dowjones closed price
- R_{FTSE} refers to FTSE index daily return and $FTSEclosed$ refers to FTSE index closed price
- R_{Gas} refers to Natural Gas price return (USD/ MMBTU) and $Gasclosed$refers to Natural Gas Henry Hub closed price
- R_{Oil} refers to Crude Oil-WTI Spot Cushing price return (USD/BBL) and $Oilclosed$refers to Crude Oil-WTI Spot Cushing closed price
- R_{Gold} refers to Gold Bullion LBM price return (USD/Troy Ounce) and $Goldclosed$ refers to Gold Bullion LBM closed price
- $R_{THB/USD}$ refers to Thai Baht to Us dollar exchange rate return and $THB/USDclosed$ refers to Thai Baht to Us Dollar closed exchange rate price

2) Higher order differences of inputs

In order to generate second order inputs, 1st order difference inputs which are stocks' returns are differenced as in (1). For example, 2nd order difference of SET price index can be calculated by

$$\boldsymbol{\Delta R_{SET_{t-1}} = R_{SET_{t-1}} - R_{SET_{t-2}}} \qquad (1)$$

In addition, by differencing the inputs repeatedly, the higher order difference inputs are generated. For example, in case of SET price index, 3rd order difference and n^{th}order difference can be calculated by the equation(2) and the equation (3) respectively.

3rd order difference SET5index

$$\boldsymbol{\Delta^2 R_{SET_{t-1}} = \Delta R_{SE0_{t-1}} - \Delta R_{SET_{t-2}}} \qquad (2)$$

n^{th}order difference SET index return

$$\boldsymbol{\Delta^{\mathrm{n}-1} R_{SET_{t-1}} = \Delta^{\mathrm{n}-2} R_{SET_{t-1}} - \Delta^{\mathrm{n}-2} R_{SET_{t-2}}} \qquad (3)$$

3) Label preprocessing

The prediction of SET index will be classified into two levels of outputs which are "Up" and "Down".

According to those levels, "Up" means the return of SET index has exceeded a certain level. In this paper, the threshold level is set at 0%, if $\ln\left(\frac{SETclosed_t}{SETclosed_{t-1}}\right) > 0\%$ the return at time t will be classified into "Up" group. On contrary, if $\ln\left(\frac{SETclosed_t}{SETclosed_{t-1}}\right) \leq 0\%$ the return at time t will be classified into "Down" group.

TABLE II. OUTPUT LABEL IN CLASSIFICATION

Classification	Criteria
Up	$\ln\left(\frac{SETclosed_t}{SETclosed_{t-1}}\right) > 0\%$
Down	$\ln\left(\frac{SETclosed_t}{SETclosed_{t-1}}\right) \leq 0\%$

4) Time frame, filtering and normalizing

The data time frame used in fitting and testing the model covered from 1/1/2002 to 10/6/2013. However, non-trading day data are filtered out, thus the model will not be applicable on non-trading day. After non-trading day data are filtered, all of the inputs data are normalized. They are ready to be fed as SVM's inputs.

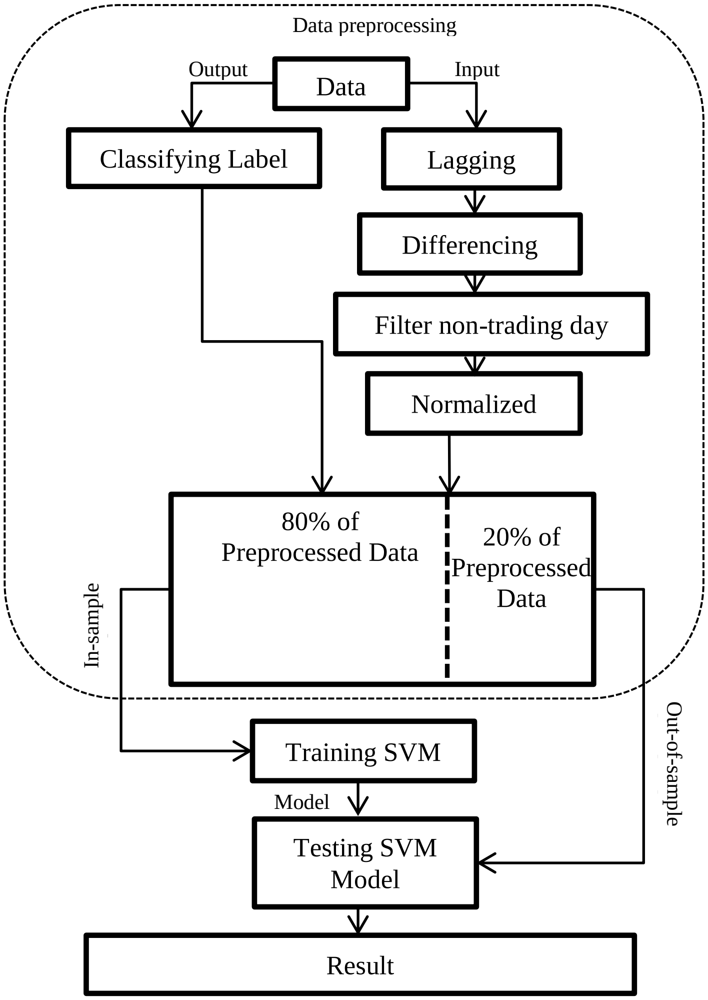

Figure 1.Flow diagram of creating the model

C. *Directional classification using the support vector machine*

The input data for the support vector machine are split into two groups which are training group (in-sample) and testing group (out-of–sample). Within the data time frame, there are 3489 samples. The first 80%, 2791 samples, will be treated as in-sample data and the remaining 20%, 698 samples, will be treated as out-of-sample data.

The in-sample dataset is applied directly to the support vector machine to fit the stock directional prediction model. After the model is fitted, the model will be applied to the out-of-sample dataset and the accuracy percentage of the out-of-sample dataset will be calculated.

Our implementation is designed as in Fig.1. and implemented in Rapidminer as shown in the Fig.2.

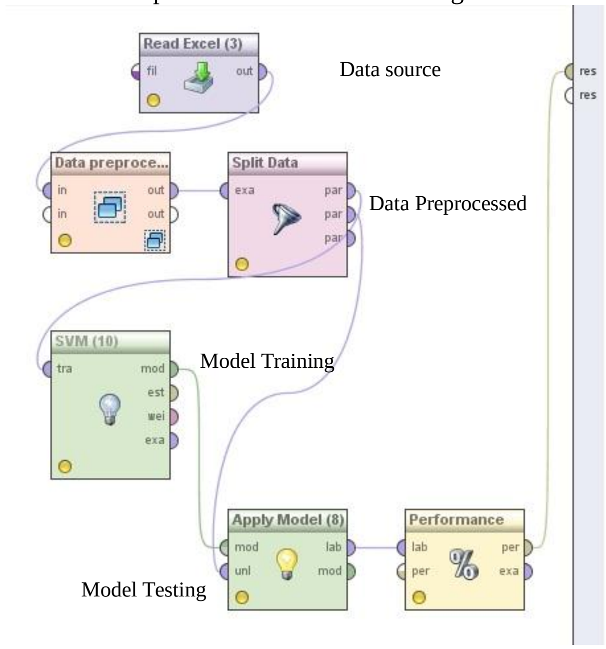

Figure 2. The overall process in Rapidminer

III. Experiment and result

A. *Support vector machine parameter settings*

In this paper, the standard SVM is used due to its popularity and simplicity in applying the model in financial prediction[14]. Also, the maximum number of iterations for the support vector machine setting is 10,000. The parameter C, which represents the tolerance of the misclassification, is set to 0.

B. *Accuracy Percentage*

To estimate the accuracy of the algorithm, the amount of the correct predictions are counted against the number of all out-of-sample data, thus accuracy percentage can be determined by

$$\textbf{Accuracy percentage} = \frac{\textbf{Amount of correct predictions}}{\textbf{Total number of data}} * \textbf{100} \quad (4)$$

C. *Directional classification result using support vector machine*

After testing the model with the 698 out-of-sample dataset when using higher order difference inputs and higher order lag inputs, the results are summarized in Table III. Table III shows the accuracy percentage at a particular number of lags and a particular number of difference orders of the input whereas “Different orders 4” means the input is including 1st, 2nd, 3rd and 4th difference orders of every input variables and “Number of lag 3” means the inputs are including 1 lag, 2 lag and 3 lag of every input variables.

TABLE III. ACCURACY PERCENTAGE OF THE SUPPORT VECTOR MACHINE CLASSIFICATION ACCORDING TO THE NUMBER OF LAGS AND DIFFERENCE ORDER OF INPUTS

		Number of lags							
		1	2	3	4	5	6	7	8
Difference orders	1	64.68%	64.07%	64.83%	65.44%	65.90%	64.37%	64.37%	65.44%
	2	64.53%	65.75%	66.06%	66.06%	64.83%	64.68%	66.06%	65.29%
	3	66.21%	66.21%	66.06%	64.83%	64.83%	64.68%	65.14%	64.68%
	4	66.36%	67.28%	65.90%	64.98%	65.14%	64.98%	64.83%	64.83%
	5	66.97%	65.75%	65.29%	65.29%	64.83%	64.68%	64.83%	64.07%
	6	66.67%	65.14%	64.53%	64.98%	64.22%	64.07%	64.07%	64.68%
	7	65.75%	64.68%	65.29%	64.22%	63.15%	64.07%	64.68%	64.37%
	8	65.29%	65.14%	65.90%	64.37%	64.22%	63.76%	64.68%	65.29%

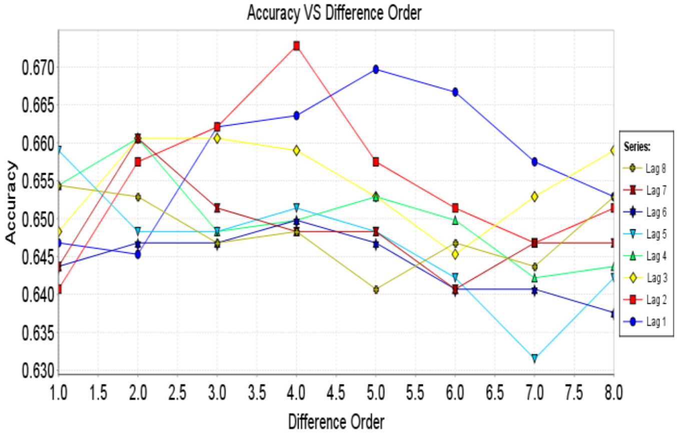

Figure 3. Accuracy percentage versus difference order of inputs at each lag

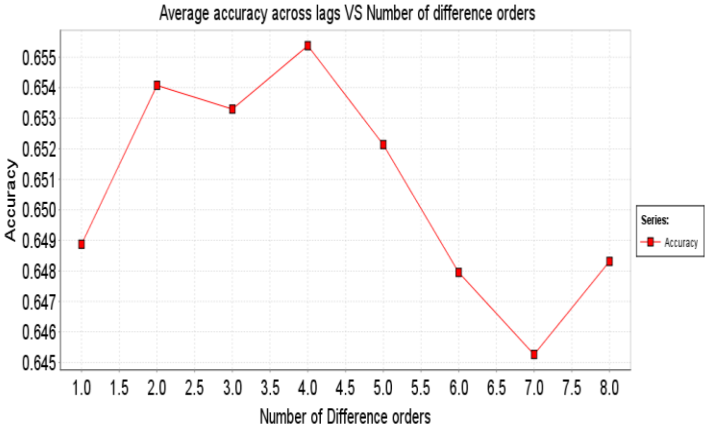

Figure 4. Average accuracy percentage along the number of lags versus the difference orders of inputs

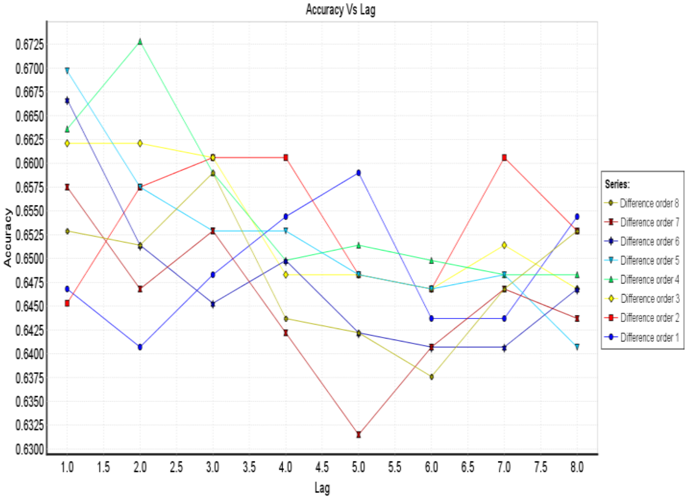

Figure 5. Accuracy percentage versus the number of lags at each the difference order

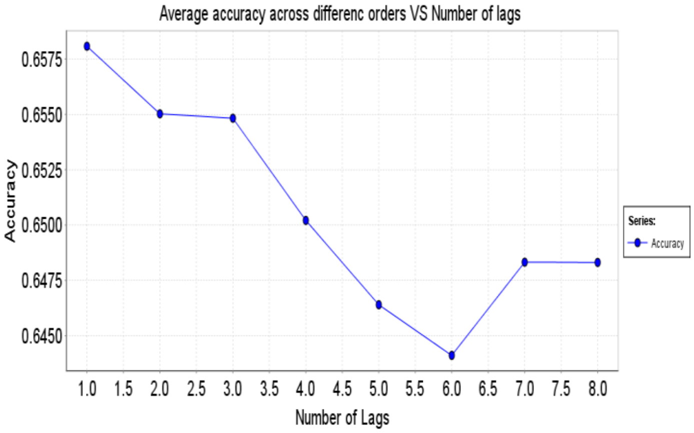

Figure 6. Average accuracy percentage along the number of difference versus the number of lags

D. *Results Analysis*

According to the result of the prediction of SET index using the support vector machine in the Fig.3, the prediction accuracy tends to increase when the different orders and lags increase, this holds for some of the particular difference orders and lags, and the prediction accuracy tends to decrease. For instance, according to Fig. 3, where number of lag is 2, the prediction accuracy is more likely to increase from 64.07% to 67.28% when the difference orders is increased from 1 to 4. The accuracy reaches its peak at 67.28% when using 4th different orders but after the different order is go beyond 4th the accuracy is more likely to decrease. In the Fig.4, the average accuracy percentage across each lag exhibits the tendency of increasing accuracy when the difference orders is less than 4 and after the difference order goes beyond 4, it shows the tendency of decreasing in accuracy.

Therefore, from these results, at low difference orders from 1 to 4, more difference orders used as inputs provide more accuracy of SVM directional prediction since differencing provides more information. This information helps predicting the directions of stock. However, at high difference orders from 5 to 7, increasing too much difference orders inputs may decrease accuracy since too high difference orders may provide useless information that overfits the prediction model.

As shown in the Fig.5 and the Fig.6, cumulating more lag as inputs may results in the drop of accuracy since more of the lags of the stock indices' return do not provide essential information in predicting stock directions.

In order to utilize the best accuracy by using this technique, additional difference orders or difference lags of inputs should be added step by step repeatedly while the accuracy is still improving. Adversely, in the case that the accuracy is dropping continuously, no more difference orders or difference lags should be added as inputs, since the model overfits the dataset.

TABLE IV. DAILY STOCK INDICES PREDICTION ACCURACY COMPARISONS

Author	**Model**	**Daily Prediction Accuracy (%)**	**Exchange Market**
Fernandez-Rodriguez et al. (2000)	NN	58	IGBM
Harvey et al. (2000)	NN	59	NYSE
Lendasse et al. (2000)	RBFN	57.2	BSE20
Francis E.H. Tay, Lijuan Cao(2001)	SVM	58.29	CME-SP
Kim K-j. (2003)	SVM	57.8313	KOSPI
Halliday (2004)	NN	55.57	NYSE
Doesken et al. (2005)	M-FIS	53.31	NYSE
Doesken et al. (2005)	TS-FIS	56	NYSE
S.-H. Hsu et al. (2009)	SVM+SOM	59.07	HSI
P. Samurwong et al. (Proposed)	SVM	67.28	SET

- NN refers to a neural network
- RBFN refers to radial basis function neural network
- SVM refers to a support vector machine
- M-FIS refers to Mamdani fuzzy system
- TS-FIS refers to Takagi-Sugeno fuzzy system
- SOM refers to self-organized map

According to the prediction accuracy from Table IV, support vector machine with higher difference orders and lags model proposed by this paper yields more outstanding directional prediction accuracy when compares with the other models, in which it reaches up to 67.28%.

IV. Conclusion

Data extraction is one of the most critical topics in the financial prediction. Adding difference orders helps improving accuracy in the directional prediction using the support vector machine, since differencing data, in some cases, provide useful information. However, adding too many differenced data in higher orders may result in lowering the accuracy, because generating and using useless information is more likely to create noise for the prediction model.

Considering in term of order of lags , lagging data tends to worsen the accuracy in the support vector machine directional prediction, since, in this case, the most recent lag data provide the most useful information and further lag seems to be unrelated with the daily prediction output.

Comparing with other models, the directional support vector machine prediction using higher order differences and lags proposed by this paper yields better result in term of accuracy. However, the prediction accuracy is comparing across various exchanges.

This paper covers the scope of the SET daily prediction using the support vector machine, hence further study of using higher order differences and order lags in the prediction can be applied with different frequencies such as weekly, monthly and quarterly as well as other exchanges including Hangseng index, NYSE, Nasdaq, etc.

Acknowledgment

We would like to thank Thaisiri Watewai, a lecturer in Financial Engineering program Chulalongkorn University, for giving advices during this study. Moreover, thank to financial engineering course, department of banking and finance, and Chulalongkorn University for technical support. Also, thank to Graduate School Chulalongkorn University for financial support; and lastly, Yotsanan Simakorn for revising the early version of this paper.

References

[1]S. Atsalakis George, P. Valavanis Kimon, Forecasting stock market short-term trends using a neuro-fuzzy based methodology, Expert Syst. Appl. 36, 2009, 10696–10707.

[2]S. Atsalakis George, P. Valavanis Kimon, Surveying stock market forecasting techniques – Part II: soft computing methods, Expert Syst. Appl. 36, 2009, 5932–5941.

Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-981-07-8859-9

[3]Doesken, B., Abraham, A., Thomas, J. & Paprzycki, M., Real stock trading using soft computing models. In Proceedings of international symposium on information technology: Coding and computing ITCC, Vol. 2, 2005, pp. 162–167.

[4]Fernandez-Rodriguez, F., Gonzalez-Martel, C., & Sosvilla-Rivebo, S., On the profitability of technical trading rules based on artificial neural networks:Evidence from the Madrid stock market. Economics Letters, 69, 2000, 89–94.

[5]M. T. Hagan, H. B. Demuth, and M. Beale, Neural Network Design. Boston, MA: PWS Publishing, 1996.

[6]Halliday, R., Equity trend prediction with neural networks. Research Letters in the Information and Mathematical Sciences, 6, 2004, 135–149.

[7]Harvey, C. R., Travens, K. E., & Costa, M. J., Forecasting emerging market returns using neural networks. Emerging Markets Quarterly, 4(2), 2000, 43–55.

[8]S.-H. Hsu, J.P.-A. Hsieh, T.-C. Chih, K.-C. Hsu, A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, Expert Syst. Appl. 36, 2009, 7947–7951.

[9]Huang, W., Nakamori, Y., & Wang, S.-Y., Forecasting stock market movement direction with support vector machine. Computer and Operations Research, 32, 2005, 2513–2522.

[10]C. Khumyoo, “The Determinants of Securities Price in the Stock Exchange of Thailand,” Master Thesis in Economics, Ramkhamhaeng University, Bangkok, Thailand, 2000.

[11]K-j. Kim, Financial time series forecasting using support vector machines. Neurocomputing, 55, 2003, 307–319.

[12]Lendasse, A., De Bodt, E., Wertz, V., & Verleysen, M., Non-linear financial timeseries forecasting application to the Bel 20 stock market index. European Journal of Economical and Social Systems, 14(1), 2000, 81–91.

[13]Lu¨tkepohl, H. and M. Kra¨tzig, Applied Time Series Econometrics, Cambridge University Press, Cambridge, 2004

[14]Perez-Cruz, F., Rodrıguez, J-A., & Giner, J., Estimating GARCH models using support vector machines. Quantitative Finance, 3, 2003, 1–10.

[15] Rapid-I, Interactive Design. Products: RapidMiner, http://rapid.com/content/view/13/69/lang.en/, 2008

[16] F. Shafait, M. Reif, C. Kofler, T.M. Breuel. Pattern Recognition Engineering. RapidMiner Community Meeting and Conference, 2010.

[17] A. J. Smola and B. Schlkopf, “A tutorial on support vector regression,”*Statistic and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[18]F.E.H. Tay, L. Cao, Application of support vector machines in -financial time series forecasting, Omega, 29, 2001, 309–317.