STOCK FORECASTING USING SUPPORT VECTOR MACHINE

LUCAS K. C. LAI, JAMES N. K. LIU

Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
E-MAIL: lucaskclai@yahoo.com, csnkliu@comp.polyu.edu.hk

Abstract:

This paper compares the performance in financial market prediction of a Neural Network approach and an approach using the regression feature of SVM. The historical values used are those of the Hang Sang Index (HSI) from 2002 to 2007 and data for January 2007 and January 2008. SVM performs well in the short term forecast.

Keywords:

Support Vector Machine; Classification; Regression and Neural Network

1. Introduction

Neural Networks (NNs) and Support Vector Machines (SVMs) [9-10] are both standard, mature machine learning approaches with applications in prediction based on times series data. NNs have been used with success in pattern classification and recognition, weather forecasting, data mining and knowledge discovery, and in time series prediction tasks such as financial market prediction. stock prices and foreign exchange forecasting [14-15]. They have shown themselves to be more accurate than other AI tools, such as Genetic Algorithms and Fuzzy Logic.

SVM is used in many machine learning tasks such as pattern recognition, object classification, and with regression analysis [8] in time series prediction in Support Vector Regression, or SVR, a methodology in which a function is estimated using observed data which in turn is used to trains the SVM. It differs from traditional time series prediction methodologies in that there is no model in the strict sense – the data drives the prediction.

SMV has been used in long term stock market forecasting. Ref. [23] used an accelerated Levenberg-Marquardt algorithm to predict the stock market series of the Jakarta Stock Indices over 10 months, achieving an RMSE of 1.96%. Ref. [16] applied SVM to forecast the price trend for a single Chinese stock although not the RMSE percentage. Ref. [17] used SVM to extract rules for the first day returns US stock market IPOs, but was accurate in only only 18% of cases. Ref. [20] claimed a profit over two months using a methodology that combined news and technical indicators. Ref. [19] used SVM to forecast the direction of stock movements which was correct 73% of the time. Ref. [21] reported the use of Support Vector Regression, or (SVR) in financial data time series prediction over a five day prediction horizon.

This paper compares the prediction performances of NN and SVM in predicting exact stock prices on the Hang Seng Index (HSI) over a five-day and a 22 day horizon. As pre-processing or input selection techniques for SVR and NN [16], we make use of the 15 days Exponential Moving Average (EMA15) and relative difference in percentage of price (RDP) RDP-5, RDP-10, RDP-15, RDP-20. We carry out experiments using the software system from [6].

The rest of this paper is organized as follows. Section II describes the method of SVR and NN Section III describes our experiments and results. Section IV offers our Conclusion and outlines future work.

2. Methods

The objective of this paper is to predict the 5 days and 22 days horizon of HSI value given the yearly values of HSI. We have downloaded the historical values of HSI from 2002 to 2007 from the financial website Yahoo. They are organized by time into as four datasets. The first two sets, for the year 2006 and the period 2002 to 2006 will be used to predict the year 2007. The third and fourthsets, for the year 2007 and the period 2003 to 2007 will be used to predict the year 2006.

978-1-4244-6527-9/10/$26.00 ©2010 IEEE

The SVR or NN models make use of are two sets of parameters. The first set contains four values Open, High, Low and Close of each trading day for the NN models. The second set contains the 15 days Exponential Moving Average (EMA15), 5 days Relative Difference in Percentage of price (RDP5), RDP10, RDP15 and RDP20 for the SVR model. Fig. 1 shows the close values for 2006 and Fig.2 shows them for 2007.

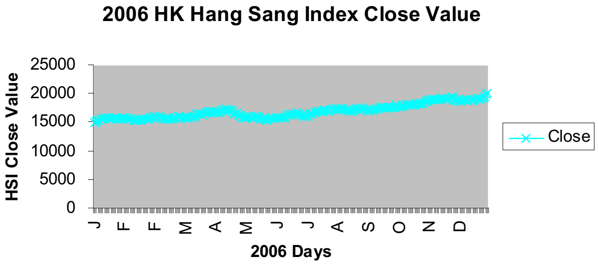

Figure 1. Data set of 2006 Hang Sang Index

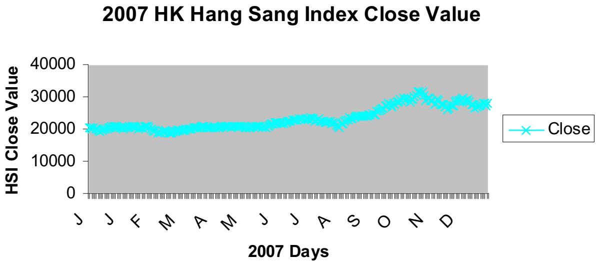

Figure 2. Data set of 2007 Hang Sang Index

The following formula is the evaluation of the predicted value.

$$\text{MAPE} = 100\frac{\sum_{i=1}^{n}\left|\frac{A-P}{A}\right|}{n}$$

MAPE stands for Mean Absolute Percentage Error which is the measure of accuracy in a fitted time series value in statistics, specifically trending. A and P are the real and the predicted values of the close value of the HSI respectively and n is the time frame or number of days. The short-term forecasting goal in this work is to predict the closing values for the first five trading days of January 2007 and January 2008. The long-term forecasting goal is to predict the entire months of January 2007 and January 2008 (22 trading days).

Given training data (x_1, y_1), ..., (x_i, y_i),, where x_i are input vectors and y_i are the associated output value of x_i, the support vector regression solves an optimization problem:

$$\min_{\varpi,b,\xi,\xi^*} \quad \frac{1}{2}\omega^T\omega + \mathrm{C}\sum_{i=1}^{l}(\xi_i+\xi_i^*) \qquad (1)$$

subject to $y_i - (\omega^T\phi(x_i)+b) \le \varepsilon + \xi_i$,

$$(\omega^T\phi(x_i)+b) - y_i \le \varepsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \ge 0,\ i=1, \ldots, l,$$

where x_i is mapped to a higher dimensional space by the function φ and ξ_i is the upper training error (ξ_i^* is the lower) subject to the ε–insensitive tube $|y - (\omega^T\phi\ (x)+b| \le \varepsilon$. The parameters which control the regression quality are the cost of error C, the width of the tube ε, and the mapping function φ.

The constraints of (1) imply that we should put most data x_i in the tube $|y - (\omega^T\phi\ (x)+b| \le \varepsilon$. If x_i is not in the tube, there is an error ξ_i or ξ_i^* which we must minimize in the objective function. SVR avoids underfitting and overfitting the training data by minimizing the training error $\mathrm{C}\sum_{i=1}^{l}(\xi_i+\xi_i^*)$ as well as the regularization term $\frac{1}{2}\omega^T\omega$. In traditional least-square regression, ξ is always zero and data are not mapped into higher dimensional spaces. Hence, SVR is a more general and flexible treatment of regression problems.

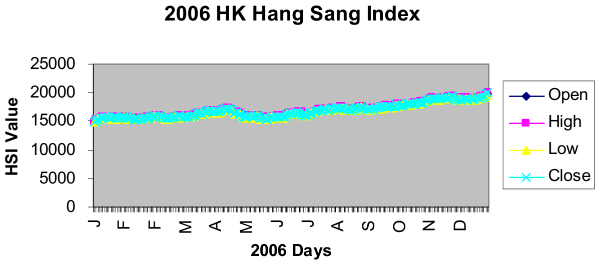

Figure 3. Similarity of the Open, High,Low and Close in 2006

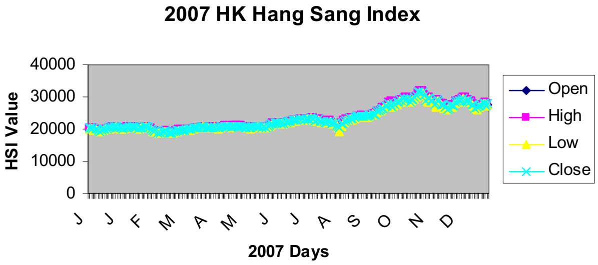

Figure 4. Similiarity of Open, High, Low and Close in 2007

Fig.3 and Fig.4 shows a trendline for Open, High, Low and Close for 2006 and 2007 datasets.

3. Experiments and results

In the following we describe two experiments, the first using an SVR software [6] and the other using Neural Networks.

a) Experiment on SVR

The type of SVM is set to 4 which is nu-SVR. The following results are from the third attributes for the SVR model (EMA15, RDP5, RDP10, RDP15 and RDP20). The cost constant as in equation (1) has different settings. The g which stands for the gamma in the kernel function also has different settings.

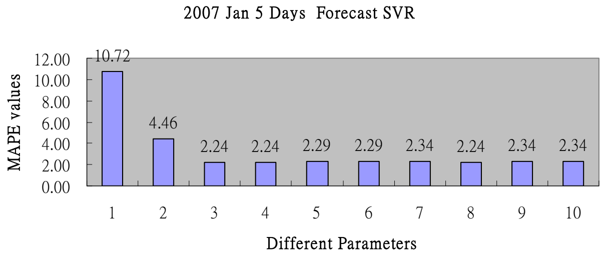

Figure 5. Prediction error for 5 days in 2007 Jan using 2006 dataset

In Fig.5, the settings for C and g are 1. C1000 g1, 2. C2000 g10, 3. C8000 g1, 4. C10000 g50, 5. C10000 g1, 6. C20000 g1, 7. C30000 g1, 8. C15000 g1, 9. C18000 g1, 10. C19000 g1, and they are displayed in the X-axis of the above figure. The best result is 2.24 and there are 3 c settings. We choose the lowest c setting which is c8000 and g1 because it is the threshold point where the change in c value has no impact on the MAPE value.

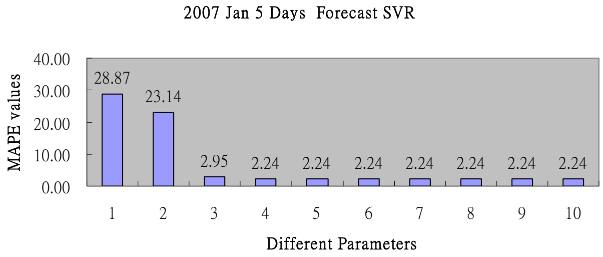

Figure 6. Prediction error for 5 days in 2007 Jan using 2002 to 2006 dataset

The x-axis of Fig. 6 shows the settings for C and g: 1. C1000 g1, 2. C2000 g10, 3. C8000 g1, 4. C10000 g50, 5. C10000 g1, 6. C20000 g1, 7. C30000 g1, 8. C15000 g1, 9. C25000 g1, 10. C50000 g1, axis. The best result is 2.24 and there are 7 c settings. We choose the lowest setting which is c10000 and g1.

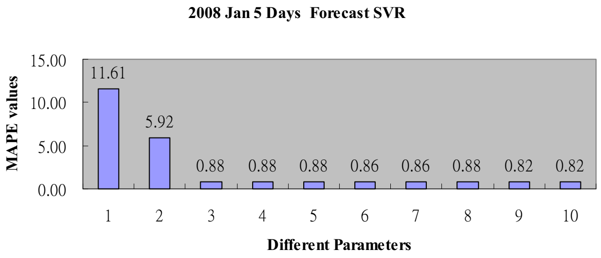

Figure 7. Prediction error for 5 days in 2008 Jan using 2007 dataset

The x-axis of Fig. 7 shows the settings for C and g: 1. C1000 g1, 2. C2000 g10, 3. C8000 g1, 4. C10000 g50, 5. C10000 g1, 6. C20000 g1, 7. C30000 g1, 8. C15000 g1, 9. C18000 g1, 10. C19000 g1. The best result is 0.82 and there are 2 c settings for c. We choose the lowest which is c18000 and g1.

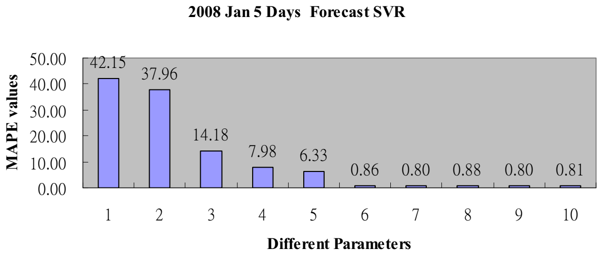

Figure 8. Prediction error for 22 days in 2008 Jan using 2003 to 2007 dataset

The x-axis of Fig. 8 shows the settings for C and g: 1. C1000 g1, 2. C2000 g10, 3. C8000 g1, 4. C10000 g50, 5. C10000 g1, 6. C20000 g1, 7. C30000 g1, 8. C15000 g1, 9. C32000 g1, 10. C28000 g1. The best result is 0.8 and the setting for c is 30000 and for g is 1.

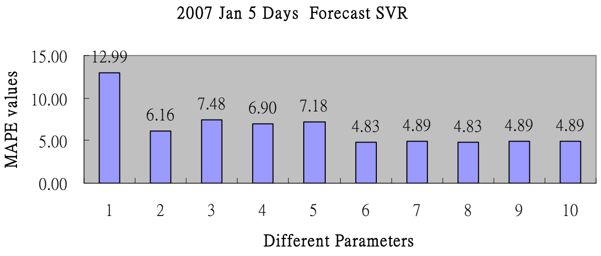

Figure 9. Prediction error for 22 days in 2007 Jan using 2006 dataset

The x-axis of Fig. 9 shows the settings for C and g: C1000 g1, 2. C2000 g10, 3. C3000 g1, 4. C4000 g1, 5. C3500 g1, 6. C80000 g1, 7. C30000 g1, 8. C15000 g1, 9. C18000 g1, 10. C19000 g1. The best result is 4.83 and the setting for c is 50000 and for g is 1.

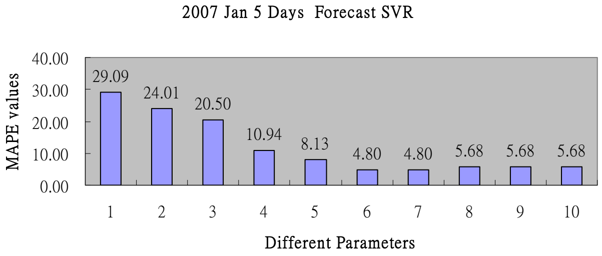

Figure 10. Prediction error for 22 days in 2007 Jan using 2002 to 2006 dataset

The x-axis of Fig. 10 shows the settings for C and g. C1000 g1, 2. C2000 g10, 3. C3000 g1, 4. C8000 g1, 5. C15000 g1, 6. C50000 g1, 7. C80000 g1, 8. C60000 g1, 9. C70000 g1, 10. C55000 g1. The best result is 4.8 and there are 2 c settings. We choose the lowest setting which is c50000 and g1.

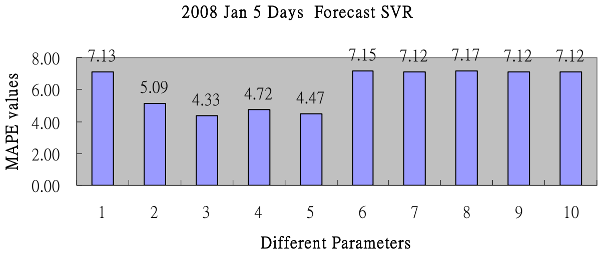

Figure 11. Prediction error for 22 days in 2008 Jan using 2007 dataset

The x-axis of Fig. 11 shows the settings for C and g. C1000 g1, 2. C2000 g1, 3. C3000 g1, 4. C4000 g1, 5. C3500 g1, 6. C80000 g1, 7. C30000 g1, 8. C15000 g1, 9. C18000 g1, 10. C19000 g1. The best result is 4.33 and the setting for c is 3000 and for g is 1.

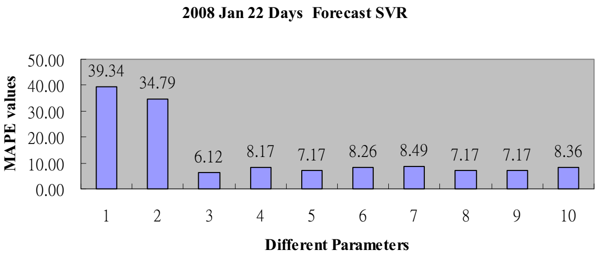

Figure 12. Prediction error for 22 days in 2008 Jan using 2003 to 2007 dataset

The x-axis of Fig. 12 shows the settings for C and g. C1000 g1, 2. C2000 g10, 3. C10000 g1, 4. C20000 g1, 5. C15000 g1, 6. C18000 g1, 7. C13000 g1, 8. C15000 g1, 9. C14000 g1, 10. C16000 g1. The best result is 6.12 and the setting for c is 10000 and for g is 1.

SVR requires effort to tune the parameters c and g in order to get a better result. It is rather difficult to determine which setting is correct. We discovered from the experiments that the parameter c has to be set in the range from 1,000 to 8000 in order to produce meaningful results and the c value must be more than 500. The best MAPE result is 0.8 for the 2008 short term forecast (5 days) which is generally better than that of the NN model. All the MAPE values are lower than 2.5. In 2001, the winning team of power loading competition by EUNITE got their MAPE as low as 1.95 [7]. In this case, the four sets of result are under 2.5 but the NN model was more accurate on the 2007 dataset.

For long term prediction (22 days), the best is 4.33 also for the 2008 dataset. The average result was 5.02 while the average for the NN model was 5.33. The NN model was more accurate on the 2007 dataset.

b) Experiments on Neural Network

We performed the NN experiments using 12 different learning functions, numbered as follows along the X-axis of Figs 13-20. 1. Hyperbolic Tangent, 2. Mixed Functions, 3. Basic, 4. Sigmoid, 5. Hyperbolic Tanh and Sine, 6. Competitive, 7. Radial Basis Function, 8. FastProp Hyperbolic Tangent, 9. FastProp Sigmoid, 10. FastProp Linear, 11. FastProp Radial Basis Function and 12. Neuro Fuzzy.. We use one hidden layerfor the 5-day and 22-day forecasts of January 2007 or January 2008 datasets. Agasin we use the Open, High, Low and Close values of each trading days. There are more than 20,000 iterations at each run. As the results for each kernel function varies between runs, we provide the results for only the best of ten runs.

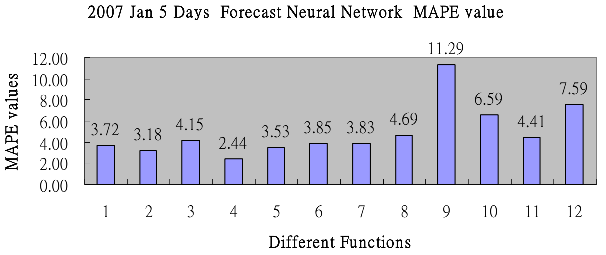

Figure 13. Prediction error for 5 days in 2007 Jan using 2006 dataset

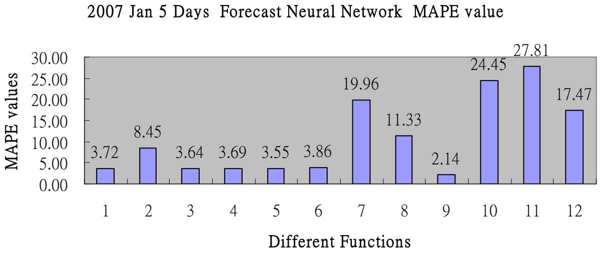

Figure 14. Prediction error for 5 days in 2007 Jan using 2002 to 2006 dataset

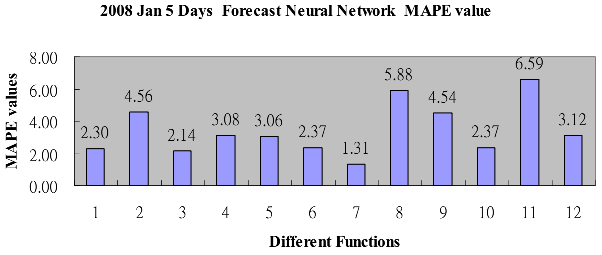

Figure 15. Prediction error for 5 days in 2008 Jan using 2007 dataset

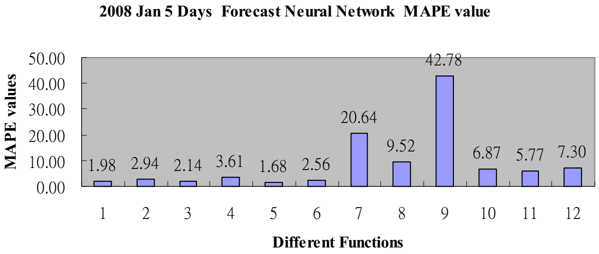

Figure 16. Prediction error for 5 days in 2008 Jan using 2003 to 2007 dataset

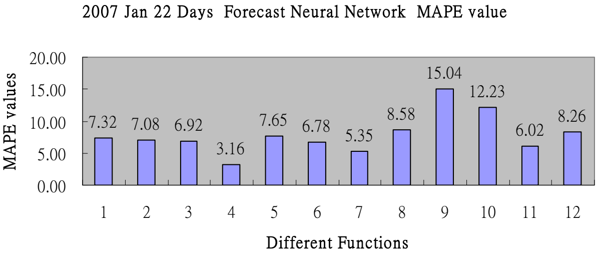

Figure 17. Prediction error for 22 days in 2007 Jan using 2006 dataset

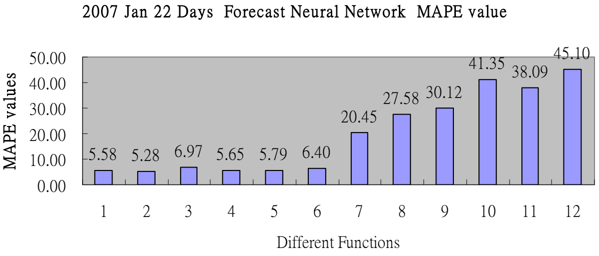

Figure 18. Prediction error for 22 days in 2007 Jan using 2002 to 2006 dataset

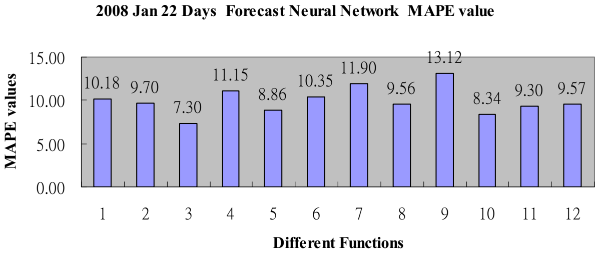

Figure 19. Prediction error for 22 days in 2008 Jan using 2007 dataset

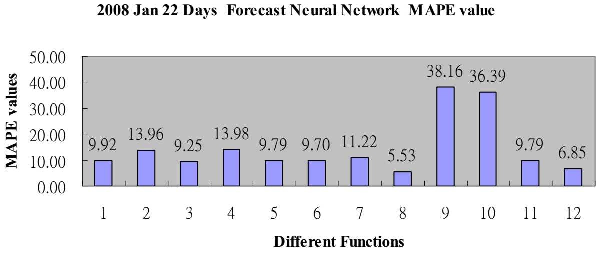

Figure 20. Prediction error for 22 days in 2008 Jan using 2003 to 2007 dataset

The following is a comparison of the NN model and the SVR model. We would like to exam the advantage and disadvantage of each model in each of the above scenario. The benchmark is the MAPE value. The lower the value the better is the performance.

In Fig.13, the best result is 2.44 and the kernel function is Sigmoid. In Fig. 5, SVR method is 2.24, which is slightly better than NN. In Fig.14, the best result is 2.14 and the kernel function is FastProp Sigmoid. In Fig.6, SVR result is 2.24, which is poor than NN. In Fig.15, the best result is 1.31 and the kernel function is RBF. In Fig.7, SVR result is 0.82, which is better than NN. In Fig.16, the best result is 1.68 and the kernel function is Hyperbolic Tanh and Sine. In Fig.8, SV result is 0.8, which is better than NN. In Fig.17, the best result is 3.16 and the kernel function is Sigmoid. In Fig. 9, SVR result is 4.83, which is poor than NN. In Fig.18, the best result is 5.28 and the kernel function is Mixed Functions. In Fig.10, SVR result is 4.8, which is better than NN but not much as both of them has value more than 5. It is not regarded as good prediction result. In Fig.19, the best result is 7.30 and the kernel function is Basic. In Fig.11, SVR result is 4.33, which is much better than NN. In Fig.20, the best result is 5.53 and the kernel function is FastProp Hyperbolic Tangent. In Fig.12, SVR result is 6.12, which is poor than NN.

The best results are produced by the hyperbolic tangent, basic, sigmoid and RBF kernels. SVR offers four standard kernel types, linear, polynomial, RBF and sigmoid and we use the default RBF for comparison. If we use sigmoid instead of the default RBF in SVR, the results are worse. As discussed in the SVR section, NN model cannot match SVR in its accuracy particular in short term forecast. In general SVR has better result in 2008 and NN has better result in 2007. We have also employed the four attributes EMA15, RDP5, RDP10, RDP15 and RDP20 to feed into the NN model for comparison but the result is even worst.

One model exceeds the other in the prediction with 2007 and 2008 data sets both for short and long term forecast. Despite the strong statistical background from SVR, it seems the advantage is not clear in this experiment. It is true from the experiment that SVR produces the best result so far. There is still room for improvement to lower the MAPE values to less than 3 in the long term forecast of SVR.

The first step in any prediction is pre-processing of the data and to find out the visible pattern before inputting into any model. However, in the 2006 and 2007 HSI datasets, it is impossible to separate the data. The 2008 dataset is more volatile than 2007 and it could be the reason why the prediction accuracy is high in long term forecast.

4. Future work and conclusion

As a conclusion, it is inconclusive to judge which model supersedes the other as the result are not supportive enough. However, it is very intriguing to point out that one model exceeds the other in the prediction with 2007 and 2008 data sets both for short and long term forecast. Despite the strong statistical background from SVR, it seems the advantage is not clear in this experiment. It is true from the experiment that SVR produces the best result so far. There is still room for improvement to lower the MAPE values to less than 3 in the long term forecast of SVR.

In future work we will attempt to improve the prediction accuracy of SVR by fine tuning the parameters c and g. To improve the accuracy of the prediction, we may use Genetic Algorithms or an advanced mathematical forecasting algorithm such as linear recurrent formulae. In pre-processing the data for each model, we have used the normalized inputs but in future may use the 12-day moving average and may include other parameters such as interest rates along with historical time series data.

Acknowledgment

The authors would like to acknowledge the partial support of the research grant of The Hong Kong Polytechnic University

References

[1] Lucas K.C. Lai and James N.K. Liu (2008), “WIPA: A Neural Network and CBR-based Model for Allocating Work in Progress”, 5th International Conference on Information Technology and Applications (ICITA 2008), Cairns Queensland Australia, 23-26 June 2008, pp. 533-538

[2] Lucas K.C. Lai and James N.K. Liu (2009), "A Neural Network and CBR-based Model for Sewing Minute Value”, in Proceedings of the IEEE 2009 International Joint Conference on Neural Networks (IJCNN 2009), Atlanta, Georgia, U.S.A 14-19 June 2009, pp. 1696-1701

[3] Lucas K.C. Lai and James N.K. Liu (2009), “ALBO: An Assembly Line Balance Optimization Model using Ant Colony Optimization”, in Proceedings of the 5th International Conference on Natural Computation (ICNC’09), Tianjin, China 14-16 August 2009, pp. 8-12 DOI:10.1109/ICNC.2009.693

[4] Lucas K.C. Lai and James N.K. Liu (2010), “WIPA: Neural Network and Case Reasoning Models for Allocating Work”, Journal of Intelligent Manufacturing, DOI:10.1007/s10845-010-0379-2 , published on line 2 February, 2010

[5] Lucas K.C. Lai and James N.K. Liu (2010), “ALBO: An Assembly Line Balance Optimization Model using Ant Colony Optimization”, Journal of Computers and Industrial Engineering, submitted 9 December, 2009

[6] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] B.J. Chen, M.W. Chang and C.J. Lin, “Load Forecasting Using Support Vector Machines: A Study on ENUNITE Competition 2001”, Department of Computer Science and Information Engineering, National Taiwan University, 2001

[8] S.H. Lee, H. Kim H. Jang and J. S. Lim, “Forecasting Short-Term KOSPI Time Series Based on NEWFM”, 2008 IEEE p 303-307 DOI:10.1109/ALPIT.2008.2

[9] Johan A. K. Suyken, Tony. V. Gestel, Jos. D. Brabanter, Bart. D. Moor and Joos Vandewalle, “Least Squares Support Vector Machines’, World Scientific Publish Co. Pte. Ltd. 2002

[10] Christopher. J.C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition”, Microsoft Research

[11] A. Majumdar, P. K. Majumdar and B. Sarkar, “An investigation on yarn engineering using artificial neural networks,” Textile Institute Vol 97 No. 5 pp.429-434.,2006

[12] C. L. Hui and S. F. Ng, “A new approach for prediction of sewing performance of fabrics in apparel manufacturing using artificial neural networks,” Textile Institute Vol 96 No. 6 pp.401-405,2006

[13] A. Wong, “Prediction of clothing sensory comfort using neural networks and fuzzy logic”, PhD thesis, The Hong Kong Polytechnic University, 2002

[14] R. Lee and J. Liu, “iJade Stock Predictor – An Intelligent Multi-Agent Based Time Series Stock Prediction System”, The Hong Kong Polytechnic University, 2001

[15] R. Lee, J. Liu and Jane You, “iJade WeatherMAN – A Multiagent Fuzzy-Neuro Network Based Weather

Prediction System", The Hong Kong Polytechnic University, 2001

[16] Y. Bao, Y. Lu and J. Zhang, "Forecasting Stock Price by SVMs Regression", Springer-Verlag Berlin Heidelberg pp. 295-303, 2004

[17] R. Mitsdorffler and J. Diederich, "Prediction of First-Day Returns of Initial Public Offering in the US Stock Market Using Rule Extraction from Support Vector Machines", Studies in Computational Intelligence (SCI) 80, 185-203 (2008)

[18] J. M. Moreira, A. M. Jorge, C. Soares and J. F. de Sousa, "Improving SVM-Linear Predictions Using CART for Example Selection", Springer-Verlag Berlin Heidelberg, pp. 632-641, 2006

[19] W. Huang, Y. Nakamori and S. Y. Wang, "Forecasting Stock market movement direction with support vector machine", DOI:10.1016/j.cor.2004.03.016 Elsevier Ltd.

[20] Y. Zhai, A. Hsu and S. K. Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction", Springer-Verlag Berlin Heidelberg,, pp. 1087-1096, 2007

[21] N. I. Sapankevych and R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey", IEEE Computational Intelligence Magazine, pp. 25-38 May 2009

[22] K. J. Kim, "Toward Global Optimization of Case-Based Reasoning Systems for Financial Forecasting", Applied Intelligence 21, 239-249, 2004

[23] F. Pasila, S. Ronni, Thiang and L. H. Wijaya, "Long-term Forecasting in Financial Stock Market using accelerated LMA on Neuro-Fuzzy structure and additional Fuzzy C-Means Clustering for optimizing the GMFs", International Joint Conference on Neural Networks, pp.3960-3965, 2008