Enhanced Online LS-SVM Using EMD Algorithm for Prices Prediction of Building Materials

Ying-Hao Yu[a], Hsiao-Che Chien[a], Pi-Hui Ting[b], Jung-Yi Jiang[a] and Pei-Yin Chen[a]*

[a] Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan
[b] Department of Business Administration, Chang Jung Christian University, Taiwan
E-mail: yinhaun@gmail.com, max6060777@hotmail.com, tphui@mail.cjcu.edu.tw, jungyi@ismp.csie.ncku.edu.tw, pychen@csie.ncku.edu.tw

Abstract -

Cost estimation is economically critical before starting off a construction project. One of the essential assignments for materials' prices prediction is to control the cost of inventory. Even though the prediction system based on support vector machine (SVM) recently has been emerged as a favourable choice, the prediction accuracy of SVM is usually deteriorated with nonstationary price data. Thus the way to explore workable price prediction still remains a challenge to be resolved for materials' cost control. In this paper, an enhanced online least squares support vector machine (LS-SVM) is proposed to predict the trend of building materials prices. Our design is to incorporate with empirical mode decomposition (EMD) to deconstruct nonlinear and nonstationary data for the set of intrinsic mode functions (IMFs), which are represented in sinusoid-like waveforms. Superior prediction, therefore, can be attained by predicting IMFs with online LS-SVMs. According to our simulation results, proposed EMD designs notably improve prediction accuracy from online LS-SVM and are workable for the cost estimation of building materials.

Keywords -

Support Vector Machine; Nonstationary Data; Empirical Mode Decomposition; Intrinsic Mode Functions; Online LS-SVM

1 Introduction

Effective cost estimation is paramount to stakeholders for evaluating viability of a construction project. Categorization of this assignment is about to control resources comprehensively of materials, labour, and equipment [1]. Of these resources, severe fluctuation of material cost from domestic and international economic influences is usually imputed for the principal cause of failure [2]. This stresses the indispensability of materials' prices prediction during pre-project planning.

In past years, the expert systems based on support vector machine (SVM) have become a main stream in statistical machine learning and prediction. The initial motivation of SVM was to classify patterns by linearly or nonlinearly separating classes using a hyperplane [3]. This idea soon had been improved for linear or nonlinear function's estimation and is also known as the support vector regression (SVR) [3], [4]. After the release of SVM and SVR, Suykens and Vandewalle proposed a refined version called least squares support vector machine (LS-SVM) [5]. In this work, least squares loss function and equality were designed to substitute for ε-insensitive loss function and inequality constraints in SVM and SVR [6]. Such reformulation significantly reduces the computing load for large data set and makes it more popular in many prediction systems [6].

Nevertheless, when support vector machine works with nonstationary data, insufficiency of this kind of system will manifest. The issue arises from SVM's identical form, which utilizes a linear function $f(x)$ involving kernel function to resolve nonlinear problems. This design sometimes is improbable in that a unique function cannot satisfy whole sequence of non-stationary time series [7]. Such limitation is inapplicable for financial analysis, e.g., exchange rate and price prediction [8].

Based on the merits of SVM, some enhancements have been developed for nonstationary systems. For instance, in the work of Chang et al. [7], single linear function of SVM is replaced by multiple functions for nonstationary time series. Zhang et al. suggest deconstructing nonstationary signal beforehand by using wavelet packet transform [9], and LS-SVM cooperating with differential evolution is designed by Chen et al [2]. For more efficiency, the algorithm of empirical mode decomposition (EMD) has been utilized to deconstruct nonstationary data for sinusoid-like signals IMFs, then SVMs can mostly predict (track) almost-stationary IMFs

[10], [11]. Another algorithm named local mean decomposition (LMD) was reported to take over EMD by better signals quality and prediction accuracy, but this algorithm is highly depending on optimal smoothening processes to signals [12].

In this paper, we propose a LSSVM-based system for prices trend prediction of building materials. Prediction accuracy of proposed system was tested with two building materials of copper and aluminium. Our system utilizes online algorithm to dynamically update training database with daily prices [13], [14]. For more accuracy, the algorithm of EMD is adopted to deconstruct nonstationary price data before online LS-SVM. Superior accuracy of online EMD LS-SVM can be further achieved by improving the prediction performance of IMF1.

This paper is arranged as the follows. In the Section 2, the algorithms of our prediction designs such as EMD, LS-SVM, and online algorithms will be detailed. Comparisons among prediction results with different LSSVM-based designs will be listed in the Section 3. A short discussion about the method to enhance online EMD LS-SVM is in the Section 4. Finally, the conclusion is drawn in the Section 5.

2 Prediction Designs

Proposed online EMD LS-SVM is shown in Figure 1. Each material's price data are firstly deconstructed by EMD algorithm for IMFs and a residual item R_n. These signals are then processed respectively by a trained online LS-SVM. The trend of material's price can then be composed by sum of online LS-SVMs' outputs. Algorithms of EMD, LS-SVM, and online algorithms are expressed in the following sections.

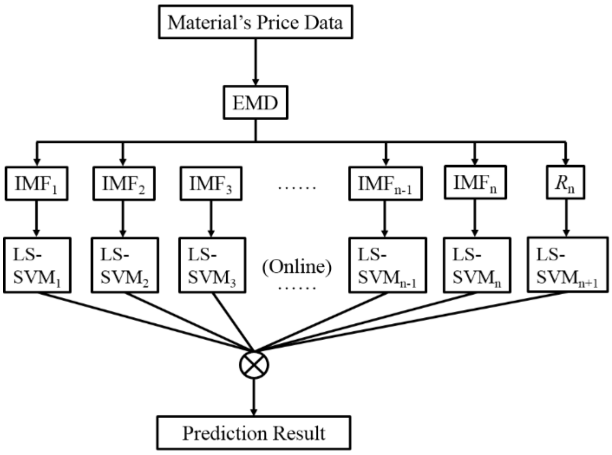


Figure 1. The configuration of proposed online EMD LS-SVM

2.1 Empirical Mode Decomposition

Empirical mode decomposition (EMD) was originally designed to deconstruct nonstationary time series data to IMFs. Huang et al. designed this algorithm for examining IMFs' instantaneous frequencies by Hilbert spectral analysis (HSA) in order to avoid complicated computing [15]. Generally, the locus of an IMF is sinusoid-like signal and similar to the harmonic of original signal. The difference is that IMF's signal might have various amplitude and frequencies [15].

The algorithm to extract IMFs can be summarized as the following steps:

1. Extract local extrema points (maxima and minima) of tested signal $x(t)$.
2. Determine upper envelope $x_u(t)$ and lower envelope $x_l(t)$ by linking up local maxima and minima respectively.
3. Derive the first mean $m_1(t)$ between upper and lower envelopes by:

$$m_1(t) = \frac{(x_u(t) + x_l(t))}{2} \tag{1}$$

4. Define the first error $h_1(t)$ as:

$$h_1(t) = x(t) - m_1(t) \tag{2}$$

where $h_1(t)$ should conform to the properties of IMF as:

a. The difference between extrema and zero-crossing points number of the whole data set must be ≤ 1.

b. The mean value of envelopes which are composed of the local maxima and minima should be zero at any point.

If $h_1(t)$ satisfies the requirements of an IMF, the first IMF function can be confirmed as $C_1(t) = h_1(t)$ — otherwise extracting procedure goes back to the step 1 and replaces $x(t)$ with $h_1(t)$.

5. Determine the residual item $R_1(t)$ as:

$$R_1(t) = x(t) - C_1(t) \tag{3}$$

6. Repeat the steps from 1 to 5 for IMF$_2$ by replacing

$x(t)$ with $R_1(t)$, and extracting processes will stop when $R_n(t)$ becomes a monotonic function.

As shown in Figure 1, the nonstationary signal can be deconstructed into IMFs and a residual function. Since EMD's outputs are the subsets of $x(t)$, original signal can be expressed as:

$$x(t) = \sum_{i=1}^{n} C_i(t) + R_n(t) \qquad (4)$$

2.2 Least Squares Support Vector Machine

Considering a given training data set which is defined as $D = \{(x_1, y_1), ...(x_l, y_l)\}$, $x_i \in R^n$, $y_i \in R$. The LS-SVM algorithm defines a linear function $f(x)$ as:

$$f(x) = \langle \omega, \varphi(x_i) \rangle + b \qquad (5)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, ω is the weight vector, b is a bias, and $\phi(x)$ represents a mapping function to map the input vectors into a high-dimensional feature space. The goal of prediction is to find a function $f(x_i)$, which has limited error to the actual targets y_i from training database. Thus equation (5) becomes an optimal problem for:

$$\min \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{l} e_i^2$$

$$subject\ \ to\ \ y_i = \langle \omega, \phi(x_i) \rangle + b + e_i, \ \ (i = 1, ..., l) \qquad (6)$$

where e_i denotes the variable of error for misclassifications, and γ is defined as the penalty parameter to minimize estimation error and maintain function's smoothness [6], [11].

To resolve equation (6), Lagrangian function can be utilized to find out ω and e. It can be written as:

$$\begin{aligned} L_{LS-SVM} = {} & \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{l} e_i^2 - \\ & \sum_{i=1}^{l} \alpha_i \{\omega \cdot \phi(x_i) + b + e_i - y_i\} \end{aligned} \qquad (7)$$

where α_i is Lagrange multiplier, which can be either in positive or negative value. The conditions for optimality of equation (7) are:

$$\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l} \alpha_i \phi(x_i) = 0 \qquad (8)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} \alpha_i = 0 \qquad (9)$$

$$\frac{\partial L}{\partial e_i} = \gamma \cdot e_i - \alpha_i = 0, \ \ (i = 1, ..., l) \qquad (10)$$

$$\frac{\partial L}{\partial \alpha_i} = \omega \cdot \phi(x_i) + b + e_i - y_i = 0, \ \ (i = 1, ..., l) \qquad (11)$$

Finally, the LS-SVM for function estimation can be re-written as:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + b \qquad (12)$$

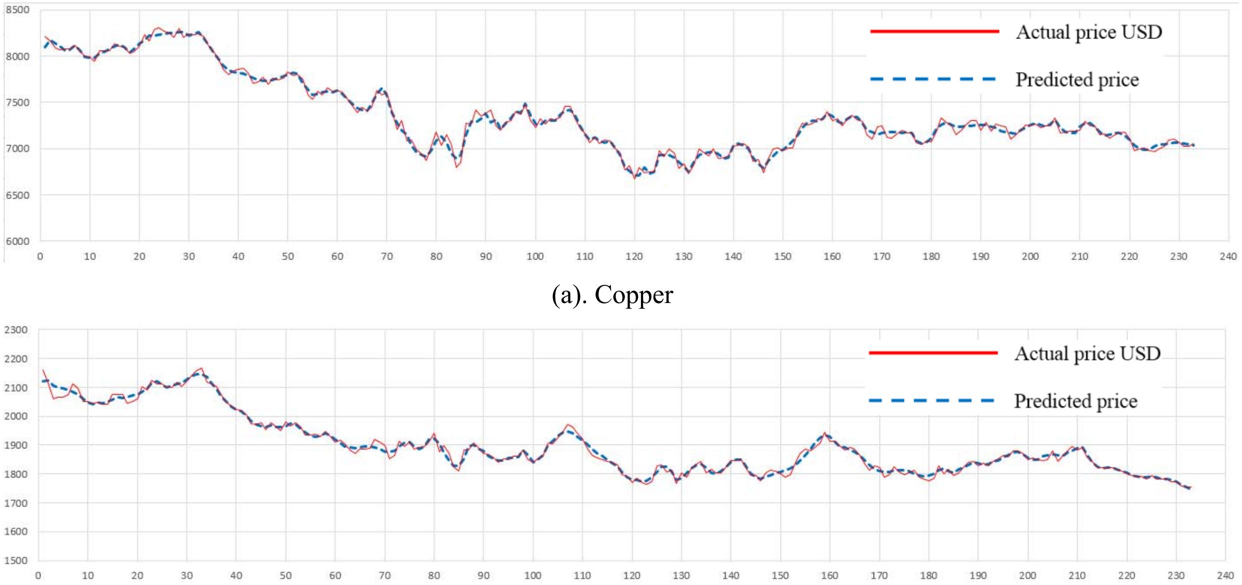
where the $K(x_i, x)$ is known as the kernel function in the form of Gaussian radial basis function (RBF) as:

$$K(x_i, x) = \exp\left(\frac{-\|x_i - x\|^2}{2\sigma^2} \right) \qquad (13)$$

The $K(x_i, x)$ must satisfy Mercer's theorem, and σ is the width of RBF. After the implementation of kernel function, we can nonlinearly map training data onto an infinite-dimensional space to resolve nonlinear problems [6], [11].

2.3 Online Algorithm for LS-SVM

Prediction accuracy of LS-SVM is depending on the features of trained data. If the features of upcoming data (signals) are different to trained data, original prediction function will lose tracking to the trend of upcoming data. This leads users confronting with such problem have to re-train system in order to learn the features of latest data. To resolve this problem, an "online" mechanism has been implemented in system to dynamically update training database during prediction. This updating mechanism can be summarized as the following steps [16]:



(a). Copper

(b). Aluminum

Figure 2. Prices predictions (ton/USD-per day) using online EMD LS-SVM

1. Add and update a training vector: After finishing prediction of a vector, a new $x_i(t)$ is added to the training database with actual value y_i.
2. Remove a vector: By following the step 1, a vector should be removed from database in order to maintain the dimension of database. However, this step might be not very urgent if the prediction system has not shortage of memory spaces and computing resources.

3 Simulation Results

Predicted building materials prices of copper and aluminium are shown in Figure 2. The data of sampled materials prices were from the futures market of London. In our test, the EMD LS-SVM were first trained with 1012 daily prices for each material and then the prices on 233 more days from different year were tested by combining with online predicting algorithms. Collected data sets of both materials in Figure .2 are obviously nonlinear and non-stationary. The Figure 2 indicates the prediction results of proposed online EMD LS-SVM can track the actual material prices efficiently.

When compare with the different prediction schemes based on LS-SVM, the criterion of mean absolute percentage error (*MAPE*) is chosen for evaluation indicator as the follows [11]:

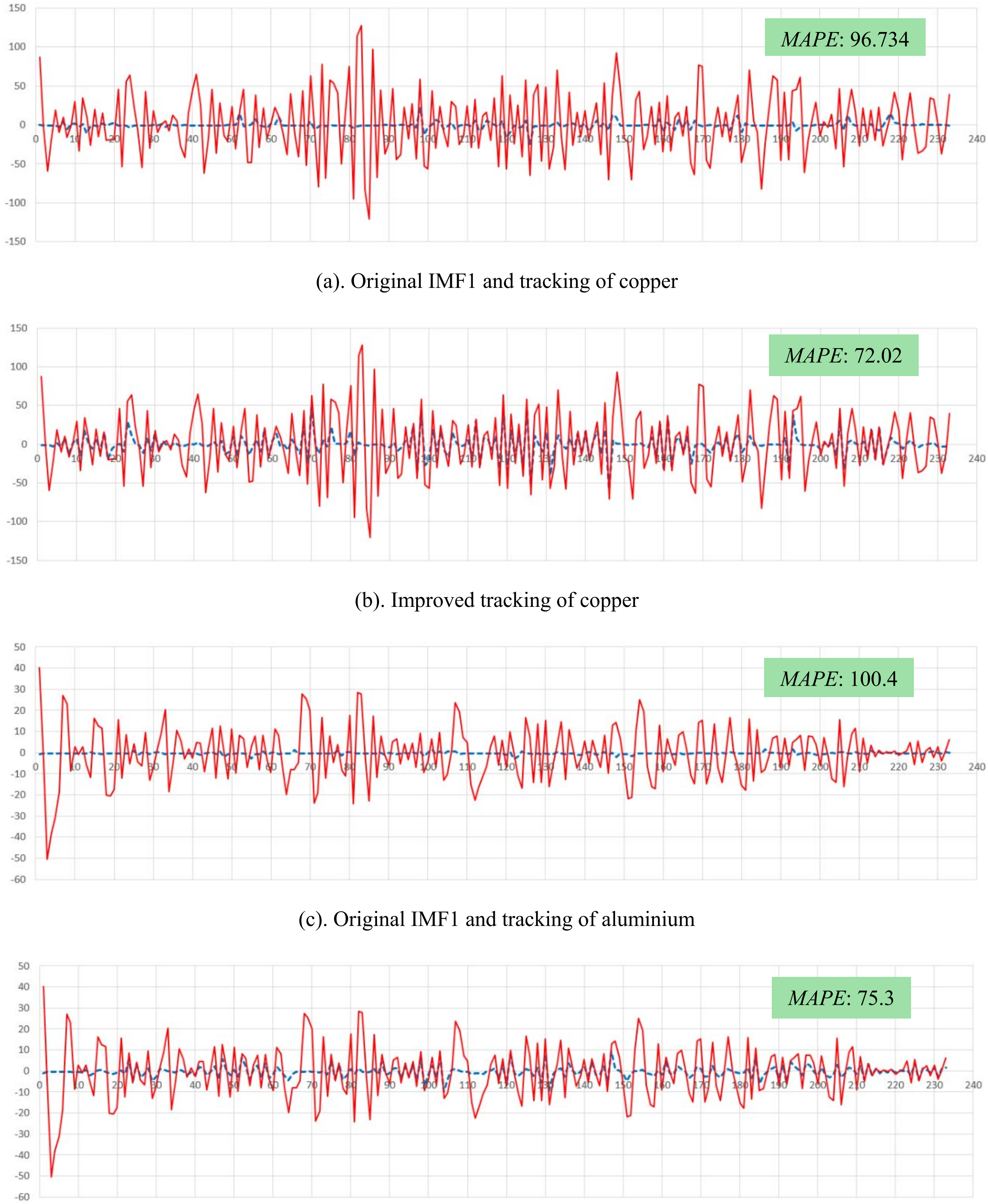
$$MAPE = \frac{1}{M}\left[\sum_{i=1}^{M}\left|\frac{r_i - f_i}{r_i}\right| \times 100\%\right] \qquad (14)$$

where r_i denotes the actual price of building material. The f_i is the predicted price, and M is the sampling number by days.

Table 1 lists prediction accuracy with different LSSVM-based designs surveyed by using *MAPE*. It can be seen that online algorithm inevitably improves the traditional LS-SVM by updating training database dynamically with incoming price data. This prediction result is even better than the EMD LS-SVM algorithm. Such a contradiction between the EMD LS-SVM and online LS-SVM arises from the sinusoid-like IMFs, which are still partially non-stationary with price data and deteriorates the tacking performance of trained LS-SVMs.

However, even though online LS-SVM has better tracking performance than LS-SVM and EMD LS-SVM for non-stationary and sinusoid-like waveforms, we can further improve online prediction accuracy if it cooperates with EMD. As shown in Table 1, the last prediction results indicates over 40% *MAPE* improvement from online LS-SVM can be achieved after incorporation of EMD algorithm.

In the end, the prediction accuracy in Figure 2 is not only sufficiently demonstrated for daily prices but also feasible for the weekly and monthly prices predictions by cooperating with online and EMD algorithms. Although we only demonstrated prediction results based on daily prices, with the same skill in the future, weekly and monthly prices can also be predicted after the accumulation of daily prices.



(a). Original IMF1 and tracking of copper

(b). Improved tracking of copper

(c). Original IMF1 and tracking of aluminium

(d). Improved tracking of aluminium

Figure 3. Comparisons between original and improved tracking for IMF1 signals

Table 1. Error comparisons among different LSSVM-based predictions

	MAPE (%)	
	Copper	**Aluminium**
LS-SVM	2.38543	6.36419
EMD LS-SVM	2.07504	4.6045
Online LS-SVM	0.87321	0.86573
EMD Online LS-SVM	0.46545	0.50844

4. Discussions

Notwithstanding EMD LSSVM-based prediction can achieve sufficient accuracy, comparisons between LMD- and EMD-LSSVM recently had been made by Dong et al. [17]. In their works, LMD-LSSVM outperformed EMD-LSSVM by lower error rate and training duration of LMD. Since EMD has been maturely developed, a question might be asked for refining EMD for better prediction performance. After surveying the signal quality of IMFs and tracking capability of online EMD LS-SVM, prediction with EMD can be further improved by manipulating on IMF1 carefully.

As shown in Figure 3(a) and 3(c), IMF1 in each test scenario is composed of high frequency transitions and is severely non-stationary comparing with the other IMFs. The reason is that IMF1 involves abrupt transitions of original data, so these signals in IMF1 caused failure of tracking by using online EMD LS-SVM. This phenomenon can be observed in Figures 3(a) and 3(c); both tracking signals coloured with dotted blue lines were nearly flatten out.

For resolving unsuccessful tracking with IMF1 signals, our answer is to mitigate the components (nonstationary parts) that cause the failure of tracking. Here we suggest improving online EMD LS-SVM's tracking by first double sampling IMF1's signal. This manipulation is achieved by interpolating a virtual price, which is from the average prices of every two days. Since the data envelopes of maxima and minima are symmetric to time axis, the average values of higher frequency transition in IMF1 will be very close to the time axis (null). In other words, by double sampling IMF1's signal, some predictions for minima-to-average or average-to-maxima will be approximated to a stationary relationship. Consequently, online LS-SVM's tracking capability with double sampling rate can then be improved. After tracking with double sampled signals, the tracked parts of virtual prices can be removed in order to restore IMF1 to original resolution but with better online LS-SVM tracking performance.

The improvements of tracking IMF1's signal by manipulation of double sampling are shown in Figures 3(b) and 3(d). By comparing with *MAPE*s before double sampling, the error rates are all significantly decreased. The remainders in IMF1 which cannot be tracked by online EMD LS-SVM are mostly with larger amplitude and could be classified as the noise. Here proposed double sampling mechanism on IMF1 makes an online EMD LS-SVM error rate 0.379% for copper and 0.423% for aluminium. Both results are also over 50% improvement from online LS-SVMs' *MAPE*, as shown in Table 1. It can be seen that proposed double sampling on IMF1 has the potentiality for a better EMD-based prediction in the future.

5. Conclusion

This paper has proposed a LSSVM-based system for prediction of building materials' prices. Higher prediction accuracy is achieved by online mechanism for dynamically updating training database. Moreover, the deconstruction of nonlinear and non-stationary data by empirical mode decomposition (EMD) can enhance prediction accuracy by tracking sinusoid-like IMFs and residual signals using online LS-SVMs. Comparing with latest LMD algorithm, double sampling IMF1's signal can provide competitive improvement of prediction accuracy. Based on our successful works on enhanced prediction of materials prices, the future work will focus on the improvement of tracking capability for IMFs and expand prediction ranges for weekly and monthly materials prices.

Acknowledgement

Supports from Ministry of Education Taiwan for the Top University Project to the National Cheng Kung University (NCKU) is gratefully acknowledged.

References

[1] Kim, H. J., Seo, Y. C., and Hyun, C. T. A hybrid conceptual cost estimating model for large building projects. *Automation in Construction*, 25:72-81, 2012.

[2] Cheng, M. Y., Hoang, N. D., Chong, T. T. Hybrid intelligence approach based on LS-SVM and Differential Evolution for construction cost index estimation: A Taiwan case study. *Automation in Construction*, 23:306-313, 2013.

[3] Vapnik, V. *The nature of statistical learning theory*, Springer, New York, 1995.

[4] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik V. Support vector regression machines. *Advanced in Neural Information Processing System*, 9:155-161, 1997.

[5] Suykens, J. A. K and Vandewalle, J. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293-300, 1999.

[6] Wang, H and Hu, D. Comparison of SVM and LS-SVM for regression. In *Proceedings of the International Conference on Neural Networks and Brain (ICNN&B '05)*, pages 279-283, Beijing, China, 2005.

[7] Chang, M. W., Lin, C. J., and Weng, R. C. Analysis of non-stationary time series using support vector machines. *Pattern Recognition with Support Vector Machines Lecture Notes in Computer Science*, Springer, 2002.

[8] Liew, V. K., Baharumshah, A. Z., and Chong, T. T. Are Asian real exchange rates stationary. *Economics Letters*, 83(3):313-316, 2004.

[9] Zhang, M., Li, K., and Hu, Y. Classification of power quality disturbances using wavelet packet energy entropy and LS-SVM. *Energy and Power Engineering*, 2:154-160, 2010.

[10] Fan, J and Tang, Y. An EMD-SVR method for non-stationary time series prediction, In *proceedings of the International Conference on Quality, Reliability, Maintenance, and Safety Engineering*, pages 1765-1770, Chengdu, China, 2013.

[11] Lin, C. S., Chiu, S. H., and Lin, T. Y. Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Economic Modelling*, 29:2583-2590, 2012.

[12] Wang, Y., He, Z and Zi, Y. A comparative study on the local mean decomposition and empirical mode decomposition and their applications to rotating machinery health diagnosis. *Journal of Vibration and Acoustics*, 132, 2010, 10 pages.

[13] Vong, C. M., Wong, P. K., Li, K., and Zhang, R. A study on online LS-SVM for modelling of electronically-controlled automotive engine performance, In *proceedings of the International Symposium on Advanced Vehicle Control*, Kobe, Japan, 2008. (6 pages)

[14] Chao, S., Zhang, J., Liu, F., and Li, M. An online LS-SVM prediction model of building space cooling load. *Advanced Materials Research*, 354-355:789-793, 2012.

[15] Huang, N. E., Shen, Z., and Long, S. R. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 454:903–995, 1998.

[16] Martin, M. On-line support vector machine regression, In *proceedings of the 13th European Conference on Machine Learning*, pages 282-294, Helsinki, Finland, 2002.

[17] Dong, Z., Tian, X., and Zeng, J. Mechanical fault diagnosis based on LMD approximate entropy and LSSVM. *TELKOMNIKA*, 11(2):803-808, 2013.