Available online at: https://ijact.in

Date of Submission	25/10/2018
Date of Acceptance	23/11/2018
Date of Publication	30/11/2018
Page numbers	2921-2925 (5 Pages)

ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

HYBRID APPROACH OF AN EMPIRICAL MODE DECOMPOSITION AND WAVELET SUPPORT VECTOR MACHINE FOR FORECASTING SINGAPORE TOURIST ARRIVALS TO MALAYSIA

A. Rafidah[1], A. Shabri[2]

[1]Technical Foundation Department, Universiti Kuala Lumpur (UniKL), Malaysian Institute of Industrial Technology, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750, Johor, Malaysia, [1]rafidahali@unikl.edu.my

[2]Departments of Mathematics, Science Faculty, University of Technology Malaysia, Skudai, Johor Malaysia

Abstract: Time series modelling and forecasting has fundamental importance to various practical domains. Thus, a lot of active research works is going on in this subject during several years. Many important models have been proposed in literature for improving the accuracy and efficiency of time series modelling and forecasting. This study presents a hybrid model empirical mode decomposition (EMD), wavelet and support vector machine (SVM) for Singapore tourist arrival to Malaysia. The EMD method is employed to decompose the monthly data tourist arrival to several intrinsic mode functions (IMFs) and residual. Wavelet support vector machine (WSVM) combines the advantages of wavelet analysis and SVM and improves the learning efficiency and forecasting accuracy. The weight of combination model is decided by forecasting precision of EMD model and WSVM model. At last, the EMD_WSVM model is used to forecast monthly data of tourist arrivals from Singapore to Malaysia and the results show that the proposed combination model has better performance on forecasting accuracy compared with the other models.

Keywords: Empirical mode decomposition (EMD); support vector machine (SVM); Wavelet support vector machine (WSVM).

I. INTRODUCTION

Time series prediction is widely applied in coal mines, electrical industry, finance, communication, tourism and many other different fields [1]. Refer [2] the time series analysis is importantly suggested on the many application including the control of physical systems, process of engineering, biochemistry, environmental economic system, forecast of arrival tourist acts on the macroeconomic policy, business management and individual decision-making establishment. Many experts and scholars recently have put forward a lot of methods based on nonlinear theories and its combinations for time series forecasting.

When we model tourist arrival time series using SVM or ARIMA, we must remember that these tourist arrival time series are inherently nonlinear and non-stationary. If we ignore this problem, it will result in worse forecasting. Therefore, hybrid models are widely used to solve the limitations in tourist arrival time series forecasting. Empirical mode decomposition (EMD) is suitable for

tourist arrival time series in terms of finding fluctuation tendency, which simplifies the forecasting task into several simple forecasting subtasks. EMD as a time–frequency resolution approach offers a new way by which the - stationary and nonlinear behaviour of time series can be decomposed into a series of valuable independent time resolutions [3]. It also can reveal the hidden patterns and trends of time series, which can effectively assist in designing forecasting models for various applications [4].
However, EMD still have limitation which this model cannot deal well with the signal decomposition effects [5]. But Wavelet analysis (WA) that proposed by Morlet et al in 1982, shown that this method performed in signal processing and can be used to decompose an observed time series (such as tourist arrival time series) into various components so that the new time series can be used as inputs for SVM models [6]. Therefore, this research will come out with new hybrid model EMD_WSVM to enhance the reliability, robustness and accuracy of separate forecasts by combining forecasts from different statistical models. The proposed approach will be compared with existing SVM approaches and WSVM hybrid models to verify whether the proposed model has more excellent features.

II. METHODOLOGY

Data Collection

In this paper, the data gather from monthly Singapore tourist arrival to Malaysia. The data was collected from January 1999 to December 2015 (204 data set). The monthly Singapore tourists were selected for this study. Fig. 1 shows the plot of the data.

Figure 1: Singapore Tourist Arrival to Malaysia January 1999 to December 2015

From Fig. 1, we can see that the data fluctuated in non-seasonal and non-stationary pattern. According to Xu et. al. [7], EMD is a process method generally for nonlinear and non-stationary data set. Thus, this paper attempts to apply a hybrid methodology which is an integration of SVM, wavelet and EMD to forecast the Singapore tourist arrival to Malaysia.

SVM Model

The SVM is a new technique for regression. The basic concept of the SVM is to map nonlinearly the original data x into higher dimensional feature space. The SVM predictor is trained using a set of time series history values as inputs and a single output as the target value. [8]. Consider a given training set of n data points $\{x_i, y_i\}_{i=1}^{n}$ with input data $x_i \in \Re^n$, n is the total number of data patterns and output $y_i \in \Re^n$. SVM approximate the function in the following form:

$$y(x) = w^T \phi(x) + b \tag{1}$$

where $\phi(x)$ represent the higher dimensional feature space, which is nonlinearly mapped the input space x. In SVM for function estimation, the estimation by minimizing regularized risk function:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m} L_\varepsilon(y_i) \tag{2}$$

is an arbitrary penalty parameter called the regularization constant.
Basically, SVM penalize $f(x_i)$ when it departures from y_i by means of a ε-insensitive loss function:

$$L_\varepsilon(y_i) = \begin{cases} 0 & \text{if} \left|f(x_i) - y_i\right| < \varepsilon \\ \left|f(x_i) - y_i\right| - \varepsilon & \text{otherwise} \end{cases} \tag{3}$$

The minimization of expression (2) is implemented by introducing the slack variable ξ_i^- and ξ_i^+. Specifically, ε-Support Vector Regression (ε-SVR) solves the following quadratic programming problem:

$$\min_{\omega, b, \xi_i^-, \xi_i^+} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\left(\xi_i^- + \xi_i^+\right) \tag{4}$$

Subject to;

$$y_i - (\omega' \phi(x_i) + b) \le \varepsilon + \xi_i^-$$

$$\left(\omega' \phi(x_i) + b\right) - y_i \le \varepsilon + \xi_i^+$$

$$\nabla i, \xi_i^- \text{ and } \xi_i^+ \ge 0$$

The solution to this minimization problem is of the form

$$f(x) = \sum_{i=1}^{m}\left(\lambda_i - \lambda_i^*\right) K(x_i, x) + b \tag{5}$$

where λ_i and λ_i^* are the Langrage multipliers associated with the constrains $y_i - (\omega' \phi(x_i) + b) \le \varepsilon + \xi_i^-$ and $\left(\omega' \phi(x_i) + b\right) - y_i \le \varepsilon + \xi_i^+$ respectively.
The kernel function can be defined as:

$$K\left(x_i, x_j\right) = \phi(x_i)' \phi(x_j) \tag{6}$$

The value of the kernel is equal to the inner product of two vectors x_i and x_j in the feature space $\phi(x_i)$ and $\phi(x_j)$
Below are the Kernal types:

- Polynomial (homogeneous) : $k(x_i, x_j) = (x_i, x_j)^d$
- Polynomial (inhomogeneous) : $k(x_i, x_j) = (x_i \cdot x_j + 1)^d$
- Gaussian Radial Basis Function : $k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$, for $\gamma > 0$ (or $\gamma = 1/2\sigma^2$
- Hyperbolic Tangent : $k(x_i, x_j) = \tanh\left(\kappa x_i \cdot x_j + C\right)$ for some $\kappa > 0$ and $C < 0$

The radial basis kernel is a popular choice in the SVM literature. [9] Therefore our computations are based on such a kernel. [10]

Wavelet Analysis

The proposed WSVM is based on wavelet analysis. The principle of wavelet analysis is to express or approximate a signal (or function) by a family of functions generated by dilations and translations of a mother wavelet as follows:

$$k_{m,n}(z) = |m|^{-1/2} k\left(\frac{z-n}{m}\right) \tag{7}$$

where m is a dilation factor n is a translation factor; and $k(z)$ is the mother wavelet, which satisfies the following condition:

$$W_k = \int_0^\infty \frac{|F(w)|^2}{|w|} dw \langle \infty \tag{8}$$

where $F(w)$ is the Fourier transform of $k(z)$. The wavelet transform of a function $g(z)$ can be expressed as

$$W_{m,n}(g) = \langle g(z), k_{m,n}(z) \rangle \tag{9}$$

where <”,”>denotes the dot product. The right-hand side of (9) means the decomposition of $g(z)$ the function on a wavelet basis $k_{m,n}(z)$, and $W_{m,n}(g)$ are the coordinates of $g(z)$ in the space spanned by. $k_{m,n}(z)$. Then the function $g(z)$ can be reconstructed as follows:

$$g(z) = \frac{1}{W_k} \int_{-\infty}^{\infty} \int_0^{\infty} \frac{1}{a^2} W_{m,n}(g) k_{m,n}(z) dm\, dn \tag{10}$$

Equation (10) can be approximated by taking the finite terms

$$\hat{g}(z) = \sum_{i=1}^{N} W_i . k_{m,n}(z) \tag{11}$$

where W_i are the reconstruction coefficients.

Empirical Mode Decomposition (EMD) Theory

Apply the EMD technique to decompose the original data into n-IMFs and a residual component $r(t)$. An IMF resulting from the EMD procedure should satisfy two conditions: Firstly, in the entire data set, the number of extreme values (maxima plus minima) and the number of zero-crossings must either be equal or differ by at most one; and secondly at any point, the mean value of the envelope, constructed by the local maxima and minima, is zero at any point. The algorithm of EMD is described as follows:

Step 1: Identify all the local extremes including maxima and minima values in time series data $z(t)$.

Step 2: Find the upper envelope $zU(t)$ and the lower envelope $zL(t)$.

Step 3: Calculate the mean value $M_1(t) = [zL(t) + zU(t)] / 2$.

Step 4: Evaluate the difference between the original time series $z(t)$ and the mean time series $M_1(t)$.

The first IMF $h_1(t)$ is defined as $h_1(t) = z(t) - M_1(t)$

Step 5: Check whether $h_1(t)$ satisfies the two conditions of an IMF property. If they are not satisfied, we repeat steps (1 till 3) of the decomposition procedure to eventually find the first IMF.

Step 6: After we obtained the first IMF, a repetition of the above steps are necessary to find the second IMF, until we reach the final time series $r(t)$ that satisfies one of the termination criteria suggesting to stop the decomposition procedure. The whole EMD is completed. The original date series can be described as the combination of n IMF components and a mean trend $r_n(t)$; that is,

$$z(t) = \sum_{j=1}^{m} h_j(t) + r_m(t) \tag{12}$$

where m is the number of IMFs, $h_j(t)$ represents IMFs and $r_m(t)$ is the final residual, which is a constant or a trend. The EMD techniques provide a multiscale analysis of the signal as a sum of orthogonal signals corresponding to different time scales and also be-taken as a filter of high pass, band pass or low pass.

The Propose Hybrid EMD_WSVM Model

Time series of Singapore tourist arrivals are known for presence of non- stationary and non- linearity in the data. Despite being capable of handling nonlinearity, the SVM model fails to handle non-stationary for short term problems [11].

A combination of EMD, wavelet and the SVM model provides an effective way to improve the prediction accuracy for nonlinear and non-stationary time series. From the previous study hybrid wavelet and hybrid EMD, it gives motivation to implementation procedure for Singapore tourist arrival prediction using hybrid EMD_WSVM model comprises three stages, which are illustrated by a flowchart as shown in Figure 2.The EMD_WSVM forecasting method procedure that involves three stages. The three stages are:

(a) Stage I – Decomposition EMD

Apply the EMD technique to decompose the original data into n-IMFs and a residual component $r(t)$.

(b) Stage II – Decomposition Wavelet and Individual Forecasting

Apply the wavelet technique to decompose each of IMFs and residual. There are four wavelet components in this analysis for each IMFs and residual. Afterwards, use SVM model to build a forecasting model for each extracted decomposed component.

(c) Stage III – Ensemble Forecasting

The forecasts of all components are aggregated using independent SVM model, which model the relationship among the EMD components and wavelet components. Then, the forecasted values all extracted SVM model are summed together to produce the final forecasting for the original time series.

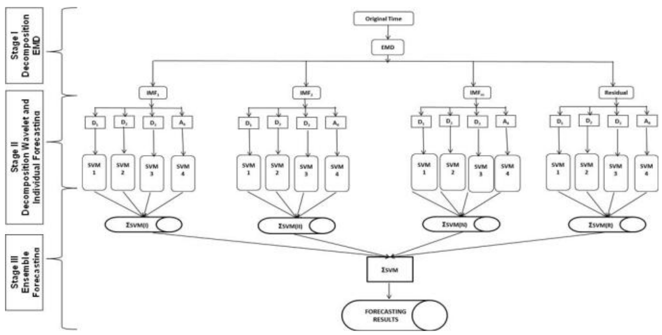

Figure 2: A flow chart of hybrid EMD_WSVM model

Forecast Performance Measures

To evaluate the forecast performance, this study use three measurement errors namely mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean squared error (RMSE). The definition are expressed as

$$\text{RMSE} = \sqrt{\frac{\sum_t^n e^2{}_t}{n}} \qquad (13)$$

$$\text{MAE} = \frac{\sum_t^n e_t}{n} \qquad (14)$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right| \text{x}\,100 \qquad (15)$$

where; n is the number of observations, t is observed time e_t is the difference between actual value (y_t) and fitted value (y_t)

TABLE I: CRITERIA OF MAPE (SOURCE: LEWIS (1982) AND HSU ETC (2009))

MAPE (%)	*Forecasting power*
<10	Excellent
10-20	Good
20-50	Reasonable
>50	Incorrect

III. RESULTS

In this section, two benchmark methods were implemented for Singapore tourist arrivals forecasting to perform a comparison with the proposed EMD_WSVM model. First is another hybrid forecasting model, one that integrates wavelet with SVM, EMD is applied to decompose the data Singapore visitor time series, and gathered components that have a monotonic function, enhancing the forecasting ability of WSVM. The others are the single SVM model was directly applied to forecast Singapore arrival to Malaysia. The modeling steps of the proposed EMD_WSVM are shown in Fig. 2. Using EMD approach in the data decomposition, the data Singapore arrival time series can be decomposed into nine independent IMFs and one residual component, respectively, as illustrate in Fig. 3. These decomposition results may enhance the model's forecasting ability in terms of divided and conquer concept [4].

Then, the decomposition forecasting variables, the independent IMFs and residual components from the previous step, apply the wavelet technique to decompose each of IMFs and residual. There are four wavelet components in this analysis for each IMFs and residual. Afterwards, use SVM model to build a forecasting model for each extracted decomposed component. For SVM kernel we employ the Gaussian RBF as the kernel function of SVM. At the ensemble forecasting step, we combine all forecasted values from the individual EMD_WSVM model to compare them with the actual Singapore tourist arrival data time series, to validate the forecasting ability of the EMD_WSVM model.

The performance evaluation of each forecasting model is based on the performance measurements forecasting models in equation (13), (14) and (15).

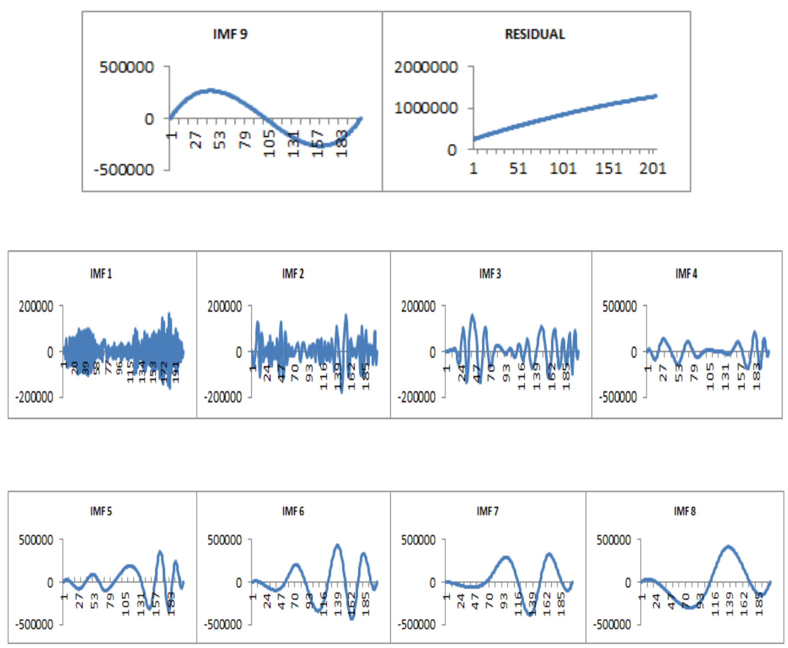

Figure 3: The IMFs and one residual for Singapore tourist arrivals data via EMD

Comparison of Forecasting Results

In order to verify the forecasting capability vof the proposed EMD_WSVM model, the WSVM and SVM models are employed for comparison, using Singapore tourist arrival data sets, RMSE,MAE and MAPE are used as performance indicators to further survey the forecasting performance of the proposed EMD_WSVM model as compared to other non linear models.

The forecasting results using EMD_WSVM, WSVM and SVM are computed and listed in Table 2, where can be seen that the RMSE,MAE and MAPE of the EMD_WSVM model are, respectively, 127591.4494, 101266.0398 and 9.02%. These values are smallest of all the forecasting models, that the deviation between the actual and forecasted values in the EMD_WSVM model is smallest. Moreover, the value MAPE proposed model refer Table 1, it showed that EMD_WSVM model is excellent model forecasting.

In sum, it can be concluded that EMD_WSVM provides better forecasting accuracy and direction criteria for Singapore tourist arrival to Malaysia than WSVM or SVM model. In addition, we see that the decomposition of time series in EMD can enhance the forecasting ability of nonlinear models. Fig. 4 showed the plot actual and forecast data the conclusion can be made is that the presented EMD-WSVM model outperforms the single SVM model and WSVM model. The outcomes gained in this research show that due to the non-linear and non stationary presence in monthly tourist arrivals data, combine EMD, wavelet and based models are more suitable for forecasting than single forecasting models. The proposed nonlinear EMD-WSVM, which is of effective decomposition and nonlinear prediction, can be used as a promising tool for time series forecasting with nonlinear and non-stationary.

Table II: The Singapore Arrival Forecasting Results

	Model Forecast		
Indicator	SVM	WSVM	EMD_WSVM
RMSE	146620.4	140065.9	127591.4494
MAE	107764.1	106283.3	101266.0398
MAPE (%)	9.55007	9.488056	9.015069033

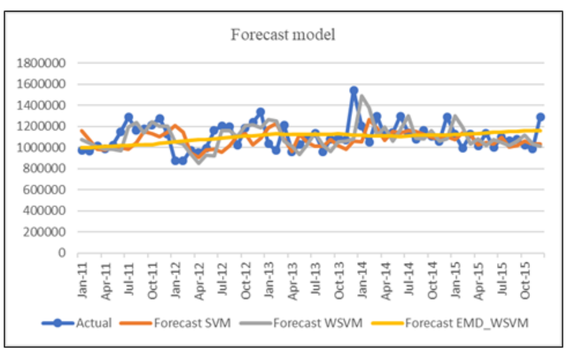

Figure 4: Forecast Model

IV. Conclusions

There has been increasing attention given to finding an effective model to address the problem of tourist arrival time series forecasting in terms of nonlinear and non-stationary characteristics. In this paper, an EMD based on wavelet and SVM forecasting model is proposed. EMD is used to detect the moving trend of tourist arrival time series data and improve the forecasting success of WSVM. Through empirical comparison of several models of Singapore tourist arrivals forecasting, the proposed model EMD_WSVM model outperforms WSVM and SVM on several criteria. Thus, it can be concluded that proposed EMD_WSVM model may be an effective tool for tourist arrivals time series forecasting.

Acknowledgement

This research was funded by a grant from University Kuala Lumpur (STRG Grant str15093).

V. References

[1] Wang, H. and Hu, D. (2005). Comparison of SVM and LS-SVM for Regression, IEEE International Conference Neural Networks and Brain; p. 279–83.

[2] Kim, G. and Yun, R.S. (2012). A Hybrid Forecast of Exchange Rate based on Discrete Grey-Markov and Grey Neural Network Model, preprint arXiv: 120.2254, Korea

[3] Tang, B., Dong, S., Song, T., (2012). Method for eliminating mode mixing of empirical mode decomposition based on the revised blind source separation. Signal Processing 92 (1), 248–258

[4] Chiun-S, L. , Sheng-H, C. and Tzu-Y, L. (2012). Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting, Volume 29, Issue 6, Pages 2583-2590

[5] Peng, L. Fan, G., Huang, M. and Hong, W. (2016). Hybridizing DEMD and Quantum PSO with SVR in Electric Load Forecasting, Energies, 9, 221; doi: 10.3390/en9030221

[6] Cheng, C.H. and Wei, L.Y. (2014). A novel time-series model based on empirical mode decomposition for forecasting TAIEX. Economic Modelling, 36:136-141. [doi:10.1016/j.econmod.2013.09.033]

[7] X. Xu, Y. Qi, & Z. Hua, Forecasting demand of commodities after natural disasters, Expert Systems with Applications, 37 (2010), 4313-4317

[8] Zhang, P. G. (2003). Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing, vol. 50, pp. 159–175

[9] Ani, S. and Suhartono (2012). Streamflow forecasting using least squares support vector machines, Hydrological Sciences Journal: vol. 57, no. 7, pp. 1275–1293

[10] Adamowski, J. and Karapataki, C. (2010). Comparison of multivariate regression and artificial neural networks for peak urban water demand forecasting: evaluation of different ANN learning algorithms, Journal of Hydrologic Engineering, vol. 15, no. 10, pp. 729–743

[11] Gupta, U., & Solanki, H. (2014). Boron: Essential Micronutrient for Plant and Animal Nutrition. International Journal of Pharmacy Research and Technology (Vol. 4, pp. 12–21).