Expert Systems with Applications 38 (2011) 10574–10578

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting

Shuhaida Ismail [a,*], Ani Shabri [a], Ruhaidah Samsudin [b]

[a] *Department of Mathematic, Science Faculty, University Technology of Malaysia, 81310 UTM Skudai, Johor, Malaysia*

[b] *Department of Software Engineering, Faculty of Computer Science and Information System, University Technology of Malaysia, 81310 UTM Skudai, Johor, Malaysia*

ARTICLE INFO

*Keywords:*
Time series
Least square support vector machine
Self-organizing maps
Forecasting

ABSTRACT

Support vector machine is a new tool from Artificial Intelligence (AI) field has been successfully applied for a wide variety of problem especially in time-series forecasting. In this paper, least square support vector machine (LSSVM) is an improved algorithm based on SVM, with the combination of self-organizing maps(SOM) also known as SOM-LSSVM is proposed for time-series forecasting. The objective of this paper is to examine the flexibility of SOM-LSSVM by comparing it with a single LSSVM model. To assess the effectiveness of SOM-LSSVM model, two well-known datasets known as the Wolf yearly sunspot data and the Monthly unemployed young women data are used in this study. The experiment shows SOM-LSSVM outperforms the single LSSVM model based on the criteria of mean absolute error (MAE) and root mean square error (RMSE). It also indicates that SOM-LSSVM provides a promising alternative technique in time-series forecasting.

## 1. Introduction

Time series analysis and forecasting is an active research area over the last few decades. Various kinds of forecasting models have been developed and researchers have relied on statistical techniques to predict time series data. The accuracy of time-series forecasting is fundamental to many decisions processes and hence the research for improving the effectiveness of forecasting models has never been stopped (Zhang, 2003). Forecasting is an important problem that spans many fields including business and industry, government, economics, environmental sciences, medicine, social science, politics, and finance. The reason that forecasting is so important is that prediction of future events is a critical input into many types of planning and decision making. In the past, conventional statistical methods were employed to forecast time series data. However, the data time series are often full of non-linearity and irregularity.

To address this, numerous artificial techniques, such as artificial neural networks (ANN) are proposed to improve the prediction result. The support vector machine (SVM) method, which was first suggested by Vapnik (1995) has recently been used in a range of applications such as in data mining, classification, regression and time-series forecasting (Zhang, 2003). The ability of SVM to solve non-linear regression estimation problems makes SVM successful in time-series forecasting.

However, Suykens, Van Gestel, De Brabanter, De Moor and Vandewalle (2002) introduced a revolution of support vector machine called least square support vector machine (LSSVM). LSSVM is modified from existed SVM. This reformulation greatly simplifies the problem in such a way that the solution is characterized by a linear system, more precisely a Karush–Kuhn–Tucker (KKT) linear system, which takes a similar form as the linear system that one solves in every iteration step by interior point methods for standard SVM.

There are various studies within the literature that used difference combining method shows that the result is better let alone using a single model. For instance, (Tay & Cao, 2001) suggest a two-stage architecture by integrating a SOM and SVR to better capture the dynamic input–output relationships inherent in the financial data. Hsu, Hsieh, Chih, & Hsu (2009) used a two-stage architecture by combining a SOM and SVR. The proposed technique is used to predict the stock price market. The result showed that the proposed technique is much better than using a single model. Kuo, An, Wang, and Chung (2006) proposed an integration using SOFM and genetic K-means. A real-world problem of the fright transport industry market segmentation is employed. The results also indicate that the proposed method is better than the other two methods used. Other than that, Kuo, Ho, and Hu (2002) integrated a SOFM and K-means for market segmentation. The simulation results indicate that the proposed scheme is slightly better

* Corresponding author. Tel./fax: +60 12 5148276.
*E-mail address:* ismail.shuhaida@gmail.com (S. Ismail).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.107

than the other conventional two-stage method with respect to the rate of misclassication, and the real-world data on the basis of Wilk's Lambda and discriminate analysis.

The main purpose of this study is to investigate the applicability and capability of time-series forecasting by comparing between SOM - LSSVM and a single LSSVM model for modeling of time-series forecasting. To verify the application of this approach, two case studies, with two datasets were used in this paper. This paper is organized as follows. In Section 2, an explanation on the basic idea of LSSVM is presented. We also covered a bit about explanation on SOM and the idea of Integration of the SOM and LSSVM. In Section 3, an experimental and result were carried out. While in section 4, we will discuss about comparison from the experiment. Finally, some conclusions are drawn. To verify the application of this approach, the benchmarked datasets are used in this study. The benchmarked datasets is well-known data sets that handled in real life time series application. There is the Wolf Yearly Sunspot Number from 1700 to 2001. The other set of data is monthly unemployed young women between ages 16 And 19 in the United States from January 1961 to August 2002.

## 2. Methodology

### 2.1. Least square support vector machine

Least square support vector machine (LSSVM) is a modification of the standard support vector machine (SVM) and was develop by Suykens et al. (2002). LS-SVM is used for the optimal control of non-linear Karush–Kuhn–Tucker systems for classification as well as regression.

Consider the first a model in the primal weight space of the following from:

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

where the $x \in \mathbb{R}^n, y \in \mathbb{R}$, and $\varphi(.) : \mathbb{R}^n \to \mathbb{R}^{nh}$ is the mapping to the high dimensional feature space. Given a sample of training set $\{x_i, y_i\}_{i=1}^{l}$ can be formulated then the following optimization problem in the primal weight space. Still combine the functional complexity and fitting error, the optimization problem of LSSVM is given as:

$$\text{Min} \qquad J(\boldsymbol{w}, \xi) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + y\frac{1}{2}\sum_{i=1}^{l}\xi_i^2, \tag{2}$$

$$\text{Such that :} \quad y_i = \boldsymbol{w}^T\varphi(\mathbf{x}_i) + b + \xi_i, \quad i = 1, 2, 3, \ldots l. \tag{3}$$

Note that this is in fact nothing else but a ridge regression cost function formulated in the feature space. However, one should be aware that when $\boldsymbol{w}$ becomes infinite dimensional, one cannot solve this primal problem. This formulation consists of equality instead of inequality constraints. Constructing the Lagrangian:

$$L(\boldsymbol{w}, b, \xi; \alpha) = J(\boldsymbol{w}, b, \xi) - \sum_{i=1}^{l}\alpha_i\{\boldsymbol{w}^T\varphi(\mathbf{x}_i) + b - y_i + \xi_i\} \tag{4}$$

where $\alpha_i \in R$ are the Langrange multipliers, which can be positive or negative in LSSVM formulation. From the optimization conditions, the following equations must be satisfied:

$$\begin{cases} \frac{\partial L}{\partial \boldsymbol{w}} = 0 \to \boldsymbol{w} = \sum\limits_{i=1}^{l}\alpha_i\varphi(\mathbf{x}_i), \\ \frac{\partial L}{\partial b} = 0 \to \sum\limits_{i=1}^{l}\alpha_i = 0, \\ \frac{\partial L}{\partial \xi_i} = 0 \to \alpha_i = \gamma\xi_i, \\ \frac{\partial L}{\partial i} = 0 \to \boldsymbol{w}^T\varphi(\mathbf{x}_i) + b - y_i + \xi_i = 0, \\ \text{For } i = 1, 2, 3, \ldots, l. \end{cases} \tag{5}$$

After elimination of the variables $\boldsymbol{w}$ and $\xi$ one obtains the following matrix solution:

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & \Omega + \frac{1}{\gamma}I \end{bmatrix}\begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}, \tag{6}$$

with $y = [y_1, \ldots, y_l]$, $1_v = [1, \ldots, l]$, $\xi = [\xi_1, \ldots, \xi_l]$, $\alpha = [\alpha_1, \ldots, \alpha_l]$, and Mercer's condition is applied within the $\Omega$ matrix:

$$\Omega_{ij} = y_i y_j \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_i) = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

The fitting function namely the output of LSSVM Regression is:

$$y(\boldsymbol{x}) = \sum_{i=1}^{l}\alpha_i K(\mathbf{x}, \mathbf{x}_j) + b, \tag{8}$$

where $\alpha_i$, $b$ are the solutions to the linear system. Although the choices of the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ in LSSVM are the same as those in SVM, more emphasis has been put on the powerful RBF kernel. Note that in the case of RBF Kernel, one has only two additional tuning parameter which is $\gamma$, $\sigma$ and $\delta^2$ as a bandwidth kernel (Suykens et al., 2002):

$$K(x, x_i) = \exp\left(-\frac{||x - x_i||^2}{\delta^2}\right). \tag{9}$$

### 2.2. Self-organizing map

The self-organizing map (SOM) which is also known as self organizing feature map (SOFM). SOM proposed by Professor Teuvo Kohonen, and is sometimes called as Kohonen map (Kohonen, 2001) is an unsupervised and competitive learning algorithm. SOM have been used widely for data analysis in some areas such as economics physics, chemistry as well as medical applications. SOM can be viewed as a clustering techniques that identifies clusters in a dataset without rigid assumptions of linearity or normality of more traditional statistical techniques (Mostafa, 2010).

The objective of SOM is to maximize the degree of similarity of patterns within a cluster, minimize the similarity of patterns belonging to different clusters, and then present the results in a lower-dimensional space. Basically, the SOM consists of two layer of artificial neurons: the input layer, which accepts the external input signals, and the output layer (also called the output map), which is usually arranged in a two dimensional structure. Every input neuron is connected to every output neuron, and each connection has a weighting value attached to it. Fig. 1 illustrates the architecture of SOM.

Output neurons will self organize to an ordered map and neurons with similar weights are placed together. They are connected to adjacent neurons by a neighborhood relation, dictating the topology of the map (Moreno, Marco, & Olmeda, 2006).

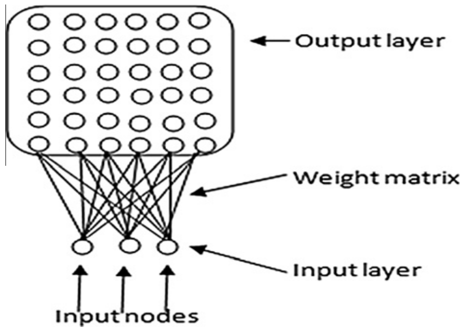

**Fig. 1.** The SOM architecture.

**Table 1**
The series data that are used to compare forecast methods.

| Series | Data | Training set | Forecasting set |
|---|---|---|---|
| A | Wolf yearly sunspot number from 1700 to 2001 | 302 | 30 |
| B | Monthly unemployed young women between ages 16 and 19 in the United States from January 1961 to August 2002 | 500 | 50 |

Given a winning neuron $I$ upon the presentation of an input $\mathbf{x}$, its updating neighborhood $\Omega_I$ starts with a wide field an gradually shrinks with time until there are no other neurons inside, i.e., $\Omega_I = \varnothing$. More specifically, we can write the updating equation for a neuron $i$ at iteration $t$ as:

$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + \eta(t)(\mathbf{x} - \mathbf{w}_i(t)), & \text{if } i \in \Omega_I(t), \\ \mathbf{w}_j(t), & \text{if } i \notin \Omega_I(t), \end{cases} \tag{10}$$

where $\eta(t)$ is the monotonically decreasing learning rate. Alternately, by using the neighborhood function $h_{Ii}(t)$, the above equation could be rewritten as:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + h_{Ii}(t)(\mathbf{x} - \mathbf{w}_i(t)). \tag{11}$$

Here, the neighborhood function is defined as:

$$h_{Ii}(t) = \begin{cases} \eta(t), & \text{if } i \in \Omega_I(t), \\ 0, & \text{if } i \notin \Omega_I(t). \end{cases} \tag{12}$$

More often, the neighborhood function takes the form of a radial basis function that is appropriate for representing the biological lateral interaction (Kohonen, 2001; Rui Xu, 2009).

### 2.3. Integrating the self-organizing map (SOM) and least square support vector machine (LSSVM)

A time series is a sequence of data points recorded sequentially in time. Time-series forecasting is used to predict future values based on past values and other variables. However, the datasets is full with non-linearity. To address of this issues, a hybrid model is employed to solve the problem. At this stage, we used a divide and conquer approach. Jacobs, Jordan, Nowlan, and Hinton (1991) were inspired by the divide-and-conquer principle that is often used to attack complex problems, i.e., dividing a complex problem into several smaller and simpler problems so that the original problem can be easily solved. A proposed model is used to predict the better forecasting result.

In the first stage, the datasets are divided into several group or cluster. In order to do this, SOM is used to cluster the training data into several disjointed clusters. Each cluster contains similar objects (Huang & Tsai, 2009). After the clustering process, an individual LSSVM model for each cluster is constructed. LSSVM can do a better forecast for each group or cluster. Meanwhile, the typical kernel functions used in this study's radial basis function (RBF) kernel (see Eq. (9)) where $\delta^2$ is the bandwidth of the RBF kernel employed some diverse kernel functions for their modeling and demonstrated that the RBF kernel has superior efficiency than other kernel. After the running an individual LSSVM for each cluster, the result will be combined in order to get the final result.

## 3. Experiment and result

### 3.1. Datasets

In this research, we examined the two well-known data sets, which is Wolf Yearly Sunspot data and monthly unemployed young women data (Wei, 2006). These data are used as a case study for this research and have been utilize the forecasting through an application aimed to handle the real life time series. These data are well known and frequently used in time series forecasting. The sunspot data are collected from year 1700 to 2001, giving a total of 332 observations. While, the unemployed data are collected from January 1961 to August 2002, giving a total of 550 observations. These data are used to demonstrate the differences between non-clustered data and clustered data using SOM.

The dataset will be divided into training set, containing the first 90% values and a test set, with the last 10%. Only the training set is used for model selection and parameter optimization, being the test set used to compare the proposed approach with other models. Information regarding the series distributed among the training and forecasting sets are given in Table 1.

### 3.2. Performance criteria

The performances of the each model for both the training data and forecasting data are evaluated. In this study, the statistical metrics is used to evaluate the result. The evaluation based on the mean absolute error (MAE) and root mean square error (RMSE), which are widely used for evaluating results of time-series forecasting. The MAE and RMSE are defined as in Table 2, where $y_t$ and $\hat{y}_t$ are the observed and the forecasted rice yields at the time $t$. The criterions to judge for the best model are relatively small of MAE and RMSE in the modeling and forecasting.

Before the training process begins, data normalization is often performed. The linear transformation formula to [0, 1] is used:

$$y_t = \frac{x_t}{x_{\max}} \tag{13}$$

where $y_t$ and $x_t$ represent the normalized and original data; and $x_{\max}$ represent the maximum values among the original data.

### 3.3. Testing the data using SOM - LSSVM

In this section, we examined the datasets using a proposal model. In order to guarantee a valid result for making predictions regarding to the new data, the dataset was randomly divided into a training set and a test set. The training set is used for model selection and parameter optimization, while the test set evaluates the prediction accuracy of the trained model.

#### 3.3.1. SOM implementation

The determination of the size of SOM is not an easy task, because the statistical properties of the data are not always available. In this study, we used Statistica 8 as a software to run the SOM. The initial map size is set at 3 × 3 units using a trail-and-error approach. The training cycle set at 1000 epochs. While a learning rate starts at 0.1 and end at 0.02, we do set the neighborhood partitions started at 3 and end with 0. One characteristic of

**Table 2**
Performance metrics and output variables.

| Performance metric | Calculation |
|---|---|
| MAE | $\frac{1}{N}\sum_{t=1}^{N}\lvert y_t - \hat{y}_t\rvert$ |
| RMSE | $\sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}$ |

**Table 3**
The result for the training and forecast using a hybrid model of SOM - LSSVM.

| Series data | Input | Training | | Forecasting | |
|---|---|---|---|---|---|
| | | MAE | RMSE | MAE | RMSE |
| A | 2 | 0.056142 | 0.071336 | 0.04222 | 0.054772 |
| | 4 | 0.043797 | 0.05742 | 0.043786 | 0.066295 |
| | 6 | 0.044466 | 0.058136 | 0.046417 | 0.062725 |
| | 8 | 0.031036 | 0.040833 | 0.04425 | 0.058988 |
| | 12 | 0.040535 | 0.053167 | 0.060639 | 0.068518 |
| B | 2 | 0.027005 | 0.034989 | 0.023153 | 0.029383 |
| | 4 | 0.028979 | 0.036853 | 0.025106 | 0.032514 |
| | 6 | 0.029542 | 0.038076 | 0.026983 | 0.032767 |
| | 8 | 0.032476 | 0.040366 | 0.03044 | 0.040346 |
| | 12 | 0.029503 | 0.038907 | 0.028978 | 0.038645 |

**Table 4**
The result for the training and forecast using a single LSSVM model.

| Series data | Input | Training | | Forecasting | |
|---|---|---|---|---|---|
| | | MAE | RMSE | MAE | RMSE |
| A | 2 | 0.0506 | 0.0692 | 0.0769 | 0.0983 |
| | 4 | 0.0402 | 0.0527 | 0.054 | 0.0712 |
| | 6 | 0.0427 | 0.0564 | 0.0563 | 0.0736 |
| | 8 | 0.0408 | 0.0538 | 0.058 | 0.0747 |
| | 12 | 0.036 | 0.0473 | 0.064 | 0.0797 |
| B | 2 | 0.0419 | 0.0533 | 0.0337 | 0.0422 |
| | 4 | 0.0337 | 0.044 | 0.0293 | 0.0394 |
| | 6 | 0.0338 | 0.0441 | 0.027 | 0.0365 |
| | 8 | 0.0329 | 0.0425 | 0.028 | 0.0375 |
| | 12 | 0.0296 | 0.0383 | 0.0286 | 0.0378 |

the SOM is that similar types of input data are mirrored to a large extent by their geographical vicinity within the representation space.

### 3.3.2. LSSVM implementation

In our experiment, we chose the Radial Basis Function (RBF) kernel where $\delta^2$ is the bandwidth of the RBF kernel as our kernel, because it tends to achieve better performance. There is no theory that can used to guide the selection of number of input. In this study the number inputs ($I$), 2, 4, 6, 8 and 12 were used for the datasets. In determining the kernel bandwidth, $\delta^2$ and the margin $\gamma$ is set at 50 and 10, respectively. The motive of choosing LSSVM is because it involves the equality constraints. Hence, the solution is obtained by solving a system of linear equations. Efficient and scalable algorithms, such as those based on conjugate gradient can be applied to solve LSSVM.

### 3.3.3. Result

The result for the training and forecast using a hybrid model of SOM - LSSVM are shown in Table 3.

By considering these training data, the lowest RMSE and MAE for series A data were calculated with input equals to eight and two for both training and forecasting. Meanwhile for B's series of datasets, the lowest RMSE and MSE for both training and forecasting were observed from input two.

## 3.4. Testing the data using single LSSVM

The selection of the number of input corresponds to the number of variables play important roles for many successful applications of this model. The issue of determining the optimal number of input is a crucial yet complicated one. In this section, we examined the datasets using a single LSSVM model only. We used the same parameters as the LSSVM's parameter for a hybrid model.

### 3.4.1. Result

The result for the training and forecast using a single LSSVM Model are shown in Table 4.

Table 4 summarizes the statistical results for training and forecasting using LSSVM models. By considering these training data, the lowest RMSE and MAE for A series data is calculated for LSSVM when no of input is twelve. For forecasting data, the lowest RMSE and MAE are calculated from four inputs. Meanwhile, the B's data was calculated as well. From the observation, the lowest RMSE and MAE was calculated from input equals to twelve, while for fore-

**Table 5**
Comparative performance between SOM - LSSVM and LSSVM For two data series.

| Data | Model | Training | | Forecasting | |
|---|---|---|---|---|---|
| | | MAE | RMSE | MAE | RMSE |
| A | SOM - LSSVM | **0.031036** | **0.040833** | **0.04425** | **0.058988** |
| | LSSVM | 0.0402 | 0.0527 | 0.054 | 0.0712 |
| B | SOM - LSSVM | **0.027005** | **0.034989** | **0.023153** | **0.029383** |
| | LSSVM | 0.0338 | 0.0441 | 0.027 | 0.0365 |

casting the lowest RMSE and MAE is come from input equals to six (see Table 4).

## 4. Comparison

For comparison purpose, the training and the forecast performance of a hybrid model SOM - LSSVM was compared with the single LSSVM model. Table 5 shows the comparison of training and forecasting precision among the two approaches based on two statistical measures for two series of datasets.

It can be observed that in the data sets for testing process, SOM - LSSVM have the smaller RMSE and MAE than a single LSSVM model in term of A series data. The results also show that the SOM - LSSVM model is outperform LSSVM for B series data as well. It means that a hybrid model of SOM - LSSVM is more competent during forecasting and training in term of RMSE and MAE.

This experiment result showed that SOM - LSSVM significantly outperformed LSSVM. The result may be attributable to the fact that SOM - LSSVM offer a better prediction. The findings in this study are compatible with the conclusions by Tay and Cao (2001).

## 5. Conclusion

This study used to compare a time-series forecasting between SOM - LSSVM and LSSVM for A and B data series. The SOM algorithm clusters the training into several disjointed cluster. After decomposing the data, LSSVM can do a better prediction. The results suggest that the two-stage architecture provides a promising alternative for time-series forecasting.

From the experimental results comparing the performance of for A and B data, it indicates that SOM - LSSVM perform better than single LSSVM. We can concluded that SOM - LSSVM provides a promising alternative technique in time-series forecasting.

## Acknowledgement

This research is supported in part by the E-Science, Ministry of Science, Technology and Innovation (MOSTI) fundamental research grant scheme under vote number 79346.

## References

Hsu, S.-H., Hsieh, J. J. P.-A., Chih, T.-C., & Hsu, K.-C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications, 36*, 7947–7951.

Huang, C.-L., & Tsai, C.-Y. (2009). A hybrid SOFM–SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications, 36*(2, Part 1), 1529–1539.

Jacobs, R. A., Jordan, M. A., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.

Kohonen, T. (2001). *Self-organizing maps*. New York: Springer; p.501.

Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. *Expert Systems with Applications, 30*, 313–324.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and *K*-means algorithm for market segmentation. *Computers & Operations Research, 29*, 1475–1493.

Moreno, D., Marco, P., & Olmeda, I. (2006). Self-organizing maps could improve the classification of Spanish mutual funds. *European Journal of Operational Research, 147*, 1039–1054.

Mostafa, M. M. (2010). Clustering the ecological footprint of nations using Kohonen's self-organizing maps. *Expert Systems with Applications, 37*, 2747–2755.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least square support vector machines*. Singapore: World Scientific.

Tay, F. E. H., & Cao, L. J. (2001). Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intelligent Data Analysis, 5*, 339–354.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Wei, W. W. S. (2006). *Time Series Analysis. Univariate and Multivariate Methods*. Pearson: New York.

Xu Rui, D. C. W. (2009). *Clustering*. IEEE.

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