HYBRIDIZING GMDH AND LEAST SQUARES SVM SUPPORT VECTOR MACHINE FOR FORECASTING TOURISM DEMAND

[1]Ruhaidah Samsudin, [1]Puteh Saad & [2]Ani Shabri

[1]Department of Software Engineering, [1]Department of Software Engineering, Faculty of Computer Science and Information System, University Teknologi Malaysia, Skudai, Johor, Malaysia.

[2]Department of Mathematics, Faculty of Science, University Teknologi Malaysia, Skudai, Johor, Malaysia.

E-mail: ruhaidah@utm.my, puteh@utm.my, ani@utm.my

ABSTRACT

In this paper, we proposed a novel hybrid group method of data handling least squares support vector machine (GLSSVM) algorithm, which combines the theory a group method of data handling (GMDH) with the least squares support vector machine (LSSVM). With the GMDH is used to determine the inputs of LSSVM method and the LSSVM model which works as time series forecasting. The aim of this study is to examine the feasibility of the hybrid model in tourism demand forecasting by comparing it with GMDH and LSSVM model. The tourist arrivals to Johor Malaysia during 1970 to 2008 were employed as the data set. The comparison of modeling results demonstrate that the hybrid model outperforms than two other nonlinear approaches GMDH and LSSVM models.

Keywords: *LS-SVM, GMDH, hybrid, tourism, time series, forecasting.*

1. INTRODUCTION

In the past few decades, the tourism industry has emerged as the fastest growing sector, and has spread widely around the world. Tourism expenditure has become an important source of economic activity, employment, tax revenue, income and foreign exchange. Therefore, every country needs to understand its international visitors and tourism receipts, to help formulate responsive policies on tourism quickly. More accurately forecasting tourism demand would facilitate for assisting managerial, operational and tactical decision making of both the private and the public sector. Therefore the selection of forecasting model is the important criteria that will influence to the forecasting accuracy.

Recently, the support vector machine (SVM) method, which was first suggested by Vapnik in 1995 [1], has recently been used in various applications such as in data mining, tourism, classification, regression and time series forecasting [2],[3],[4]. The ability of SVM to solve nonlinear regression estimation problems makes SVM successful in time series forecasting. It has become a hot topic of intensive study due to its successful application in classification and regression tasks. However, the major drawback of SVM its higher computational burden because of the required constrained optimization programming. Recently, many SVM algorithms have been discussed, such as LIBSVM [5]), BSVM [6], and LSSVM [7]). Extensive empirical studies [8]) have shown that LSSVM is comparable to SVM in terms of generalization performance. LSSVM is an improved algorithm based on SVM.

Another efficient modeling technique is the GMDH algorithm. The GMDH algorithm has been successfully used to deal with uncertainty, linear or nonlinearity of systems in a wide range of disciplines such as economy, ecology, medical diagnostics, signal processing, fossil power plant process, electric power industry and control systems [9],[10],[11], [12],[13]. Some simplified approximations, such as the two-direction regressive GMDH [14]) and the revised GMDH algorithms [15],[16]), have been introduced to model dynamic systems in flood forecast and petroleum resource prediction with some success.

Several hybrid models involved GMDH algorithm are introduced and modified. More recently, combing GMDH-type with neural networks (NN) were explored for modeling highly nonlinear dynamic system Simulation results of nonlinear time series show that the proposed method is much more accurate than the GMDH and NN algorithm.

In this paper, a novel hybrid GMDH-type algorithm is proposed by integrating simple GMDH with LSSVM. The hybrid model combines GMDH and LSSVM into one methodology. The main purpose of this study is to investigate the applicability and capability of the new hybrid model for time series forecasting. To verify the application of this approach, the tourist arrival data in the state of Johor, Malaysia is chosen as the case study. Three modeling techniques, GMDH and SVM are used to generate the single model and hybrid methods namely GLSSVM are applied in order to explore the efficiency of combining forecasts in forecasting tourist arrival.

2. METHOD

2.1 LS-SVM for time series prediction

LSSVM is a new technique for regression. The LSSVM predictor is trained using a set of time series historic values as inputs and a single output as the target value. In the following, we briefly introduce LSSVM, which can be used for time series forecasting.

Consider a given training set of *n* data points $\{x_i, y_i\}_{i=1}^{n}$ with input data $x_i \in R^n$, *p* is the total number of data patterns) and output $y_i \in R$. SVM approximate the function in the following form

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

where $\phi(x)$ represents the high dimensional feature spaces, which is nonlinearly mapped from the input space *x*. In LSSVM for function estimation the following optimization problem is formulated:

$$\min J(w,e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_i^2 \tag{2}$$

Subject to the equality constraints

$$y(x) = w^T \phi(x_i) + b + e_i\,, \quad i = 1, 2, ..., n$$

The solution is obtained after constructing the Lagrange

$$L(w,b,e,\alpha) = J(w,e) - \sum_{i=1}^{N} \alpha_i \{ w^T \phi(x_i) + b + e_i - y_i \} \tag{3}$$

With Lagrange multipliers α_i. The conditions for optimality are

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \phi(x_i)\,,$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0\,,$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma\, e_i\,,$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T \phi(x_i) + b + e_i - y_i = 0\,,$$

for $i = 1, 2, \ldots, n$. After elimination of e_i and w the solution is given by the following set of linear equations:

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \phi(x_i)^T \phi(x_l) + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{4}$$

Where $y = [y_1; ...; y_n]$, $\mathbf{1} = [1; ...; 1]$, $\alpha = [\alpha_1; ...; \alpha_n]$. According to Mercer condition, the kernel function can be defined as

$$K(x_i, x_j) = \phi(x_i)^T \; \phi(x_j)\,, \qquad i,\ j = 1, 2, \ldots, n \tag{5}$$

This finally leads to the following LSSVM model for function estimation:

$$y(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \tag{6}$$

where α_i, *b* are the solution to the linear system. The RBF kernel $K(x_i, x_j) = \exp(-| x_i - x_j |^2 / \sigma^2)$ is used in this work. There are only two parameters to be tuned: the kernel setting and γ, which are selected using cross-validation.

2.2 Group of Method Data Handling(GMDH)

The algorithm of Group Method of Data Handling (GMDH) was first developed by Madala and Ivakhnenko in 1994,[17] as a multivariate analysis method for modeling and identification of complex systems. The GMDH method was originally formulated to solve for higher order regression polynomials specially for solving modeling and classification problem. General connection between inputs and output variables can be expressed by a complicated polynomial series in the form of the Volterra series, known as the Kolmogorov-Gabor polynomial):

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + ..$$

In this case, x represents the input to the system, n is the number of inputs and a are coefficients or weights. However, for most application the quadratic form are called as partial descriptions (PD) for only two variables is used in the form

$$y^{GMDH} = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \qquad (7)$$

to predict the output. To obtain the value of the coefficients a for each m models, a system of Gauss normal equations is solved. The coefficient a_i of nodes in each layer are expressed in the form

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{Y} = [y_1\ y_2 ... y_M]^T$, $\mathbf{A} = [a_0, a_1, a_2, a_3, a_4, a_5]$,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2q}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{Mp} & x_{Mq} & x_{Mp}x_{Mq} & x_{Mp}^2 & x_{Mq}^2 \end{bmatrix}$$

and M is the number of observations in the training set.

The basic steps involved in the conventional GMDH modeling by Nariman-Zadeh in 2002,[18] are follows:

Step 1: First N observations of regression-type data are taken. The collected load data are first normalized with respect to their individual base value in order to restrict the variation of data within the same level. Those normalized data are denoted by $\{x_1, x_2, \ldots, x_M\}$ where M is the total number of input. These observations are divided into two sets: the training set and testing set.

Step 2: Select ${}^{M}C_2 = M!/(M-2)!2!$ new input variables according to all possibilities of connection by each pair of inputs in the layer. Construct the regression polynomial for this layer by forming the quadratic expression which approximates the output y in equation (7).

Step 3: Identify the single best input variable out of these of these ${}^{M}C_2$ input variables, $x' = \hat{y}_i$ according to the value of mean square error (MSE). The MSE is defined by

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2$$

The input of variables that give the best results in the first layer, are allowed to form second layer candidate model of the equation (7). Set the new input $\{x_1, x_2, \ldots, x_m, x'\}$ and $M = M + 1$. Models of the second layer are evaluated for compliance by using MSE, and again the input variables that give best results will proceed to form third layer candidate models. This procedure is carried out as long as the MSE for the test data set decrease compared with the value obtained at the previous one. After the best models of each layer have been selected, the output model is selected by the MSE. The model with the minimum value of the MSE is selected as the output model.

2.3. Hybrid Model

In the proposed method, the GMDH and LS-SVM are applied to enhance the capability of hybrid model. As input variables are selected by the decision made by GMDH and LSSVM model is used as time series forecasting. The hybrid model procedure is carried out in the following step:

Step 1 : The normalized data are separated into the training and testing sets data.

Step 2 : All combinations of two input variables (x_i, x_j) are generated in each layer. The number of input variables are ${}^{M}C_2 = \frac{M!}{(M-2)!2!}$. Construct the regression polynomial for this layer by forming the quadratic expression which approximates the output y in equation (7). The coefficient vector of the PD is determined by the least square estimation approach.

Step 3 : Determine new input variables for the next layer. The output x' variable which give the smallest of mean square error (MSE) for the train data set is selected as the new input $\{x_1, x_2, \ldots, x_M, x'\}$ with $M = M + 1$.

Step 4 : In the output layer, the new input $\{x_1, x_2, \ldots, x_M, x'\}$ of the neurons in the hidden layers are use as input for the LSSVM model. The LSSVM model with the minimum value of the MSE is selected as the output model.

Step 5 : The lowest value of MSE using LSSVM model for the test data set at each layer obtained during this iteration is compared with the smallest value obtained at the previous one. If an improvement is achieved, one goes back and repeats step 2 to 4, otherwise the iterations terminate and a realization of the network has been completed. Once the final layer has been determined, only the one node characterized by the best performance is selected as the output node. The remaining nodes in that layer are discarded. And finally, the GLSSVM model is obtained.

3. EXPERIMENT

3.1 Description of Data

In order to validate the GLSVM forecasting model for tourist arrival in the state of Johor,Malaysia, the data were collected from Johor Tourism Action Council Johor, Malaysia. These data is ranging from January, 1999 to December 2008. The time series plot is given in Figure 1.

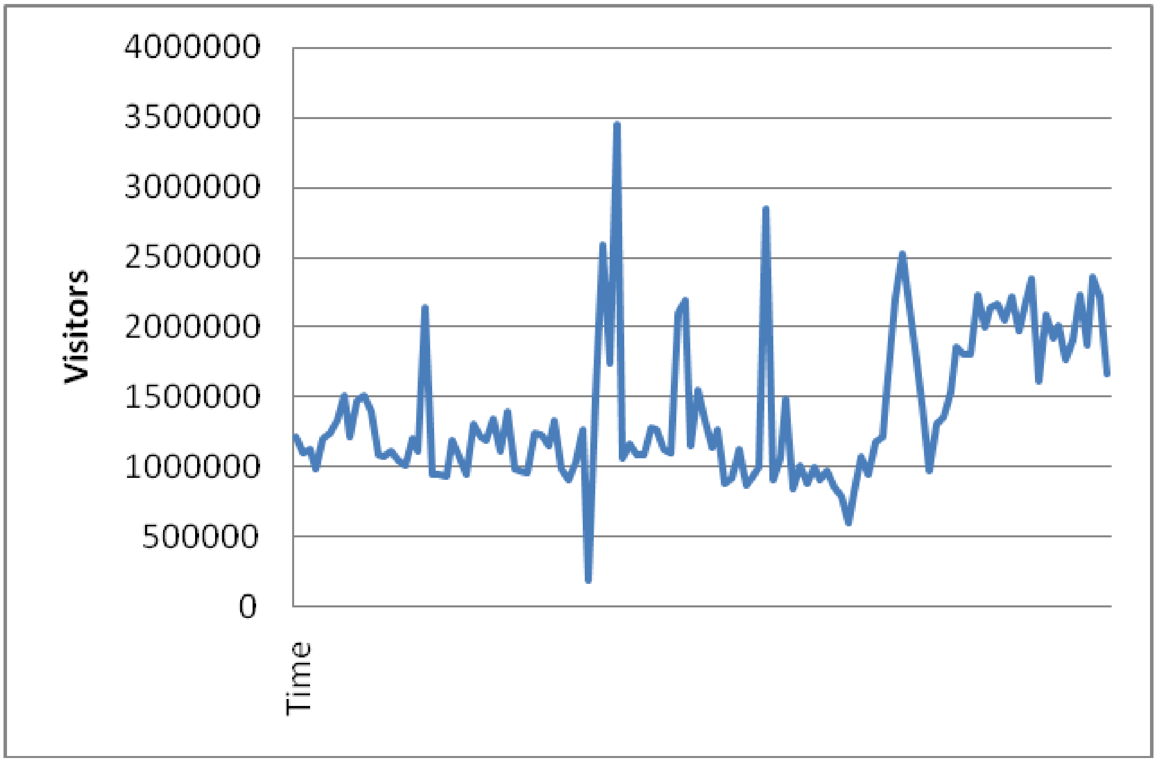

Figure 1: Visitors arrival in Johor (1999-2008)

To assess the forecasting performance of different models, each data set is divided into two samples. The first series was used for training the network (modeling the time series) and the remaining were used for testing the performance of the trained network (forecasting). We take the data from 1999 to 2008 producing 108 observations for training purpose and the remainder as the output sample data set with 12 observations for forecasting purpose.

The performances of the each model for both the training data and forecasting data are evaluated according to the mean-square error (MSE) and mean absolute error (MAE), which they are widely used for evaluating results of time series forecasting. The MAE and MSE are defined as

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2$$

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

where y_t and $\hat{y}_t$ are the observed and the forecasted rice yields at the time *t*, N equals the number of observations The criterions to judge for the best model are relatively small of MSE and MAE in the modeling and forecasting.

3.2. Results and discussion

GMDH works by building successive layers with complex connections that are created by using second-order polynomial function. The first layer created is made by computing regressions of the input variables and then choosing the best ones. The second layer is created by computing regressions of the best value in the first layer along with the input variables. Again, only the best are chosen by the algorithm. This process continues until a pre-specified selection criterion is met. In designing the GMDH model, one must determine the following variables: the number of input nodes and the number of layers. The selection of the number of input corresponds to the number of variables play important roles for many successful applications of GMDH.

The issue of determining the optimal number of input nodes is a crucial yet complicated one. There is no theory that can used to guide the selection the number of input and the number of hidden layers. To make the GMDH and the hybrid model simple and reduce some computational burden, only six levels of input nodes 2 4, 6, …, 12, and five hidden layers from 1 to 5 are selected for experiment.

There is trade-off in amount of data that can be used as input to LSSVM. The proposed architecture is trained by using LSSVM with 2 to 12 inputs. In order to obtain the optimal model parameters of the LSSVM, as was mentioned earlier, a grid search algorithm was employed in the parameter space. LS-SVM model and the grid search algorithm [19] were performed by using the MATLAB software version 8.

The precision and convergence of LSSVM is also affected by $(\gamma,\ \sigma^2)$. Cross-validation is a popular technique for estimating generalization performance. To get good generalization ability, we conduct a validation process to decide parameters. In order to better evaluate the performance of the proposed approach, we consider a grid search of (γ,σ^2) with γ in the range 10 to 1000 and σ^2 in the range 0.01 to 1.0. For each hyperparameter pair (γ,σ^2) in the search space, 5-fold cross validation on the training set is performed to predict the prediction error. Thus, a robust model was obtained by selecting those optimal parameters that give the lowest MSE in the subarea.

The proposed hybrid learning architecture is composed of two stages. In the first layer, GMDH is used to determine the inputs of LSSVM method. The second layer consists of LSSVM which work as time series forecasting. The hybrid model algorithm that is proposed in this paper is constructed in exactly the same manner as the standard GMDH algorithm except the new input variables is used for the LSSVM model.

Table 1 shows the final results of modelling and forecasting precision among the three different approaches based on RMSE and MAE. In Table 1, the lowest RMSE and MAE are found with the SVM in training and the hybrid model in forecasting. The results demonstrating that the hybrid model provides a better approach for forecasting the tourism time series.

Table 1: Comparison of GMDH, SVM and GLSSVM in training and forecasting

		GMDH	SVM	Hybrid
Training	RMSE	0.0114	0.0065	0.0062
	MAE	0.0687	0.0491	0.0478
Forecasting	RMSE	0.0051	0.005	0.0044
	MAE	0.0584	0.0554	0.0545

Figure 2 illustrates the comparison of errors between generated and observed values in the forecasting data domain. Similar to the performance in the forecasting process, the modelled error line generated from hybrid is much closer to the zero-value-line and flatter than the error lines from GMDH and LSSVM.

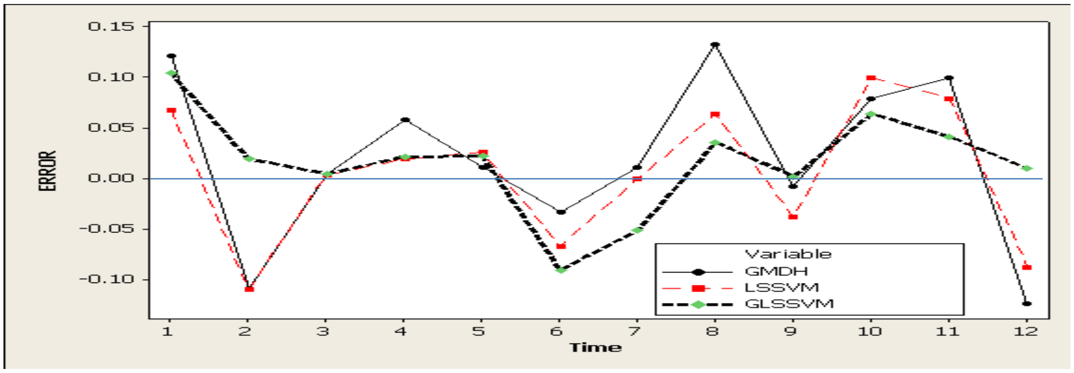

Figure 2: Comparison of errors of predictions (differences from the observed and tested values)

4. CONCLUSION

This study applied a novel hybrid model GLSSVM to the forecasting fields of tourism time series. The hybrid model of algorithm developed here combines the GMDH theory with the LSSVM approach. With the GMDH is used to determine the inputs of LSSVM method and the LSSVM algorithm which work as time series forecasting. The case study on the tourism time series data testing demonstrated that the fitting accuracy of hybrid model algorithm is superior to those of the GMDH and LSSVM. Generally, the hybrid algorithm is robust in the forecasting of nonlinear time series.

5. REFERENCES:

[1] V.N. Vapnik, "*The nature of statistical learning theory",* Springer-Verlag, New York, 1995.

[2] P.F. Pai, W.C. Hong and C.S. Lin, Forecasting tourism demand using a multifactor support vector machine model, *Springer Verlag*, Berlin Heidelberg,2005, 512-519.

[3] K.Y. Chen and C.H.Wang, "Support vector regression with genetic algorithms in forecasting tourism demand", *Tourism Management 28*,2007, 215-226.

[4] K. Zhao, Y.J. Tian & N.Y. Deng, "Robust unsupervised and semisupervised bounded c-support vector machines", 7th. IEEE *International Conference on Data Mining Workshops*, October 28-31, IEEE Computer Society, Omaha, NE, USA, 331-336.

[5] C.C. Chang and C.J. Lin," LIBSVM: A Library for support vector machines", *Department of Computer Science and Information Engineering*, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

[6] C.W. Hsu and C.J. Lin, "A comparison of methods for multi-class support vector machines", *IEEE Trans. Neural Networks, 13*, 2002, 415-425.

[7] J.A.K. Suykens, L. Lukas, P.V. Dooren, B. de Moor and J. Vandewalle, "Least Squres Suport Vector Machine Classifiers: A Large scale algorithm", *Neural Processing Letters 9*, 1999, 293-300.

[8] T. V. Gestel, J.A.K. Suykens, B Baesens, S. Viaene and J. Vanthienen , "Benchmarking least squares support vector machine classifiers", *Machine Learning*, *54*, 2004,5-32.

[9] A.G. Ivakheneko and G.A. Ivakheneko. "A Review of Problems Solved by Algorithms of the GMDH". *Pattern Recognition and Image Analysis*,1995, 5(4): 527-535.

[10] Onwubolu, G.C. P. Buryan and F. Lemke. "Modeling Tool Wear in End-Miling Using Enhanced GMDH Learning Networks", *International Journal Advance Manufacture Technology*, 2007.

[11] T. Kondo, A.S. Pandya and H. Nagashino. "GMDH-type neural network algorithm with a feedback loop for structural identification of RBF neural network". *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 2007, 11:157-168.

[12] V. Puig, M. Witczak, F. Nejjari, J. Quevedo, and J. Korbicz. "A GMDH neural network-based approach to passive robust fault detection using a constraint satisfaction backward test",*Engineering Applications of Artificial Intelligence*, 20, 2007, 886-897.

[13] F. Li, B.R. Upadhyaya and L.A. Coffey. "Model-based monitoring and fault diagnosis of fossil power plant process units using Group Method of Data Handling",*ISA Transactions*, 2009, 2: 213-219.

[14] D. Liu, and X.F. Wang. "Two-Direction Iterative Regression Algorithm of GMDH", *Information and Control*,1989, 5: 47-51.

[15] T. Kondo, "NonLinear pattern Identification by Multi-layered GMDH-Type Neural Network Self-selecting Optimum Neural Network Architecture", *Springer-Verlag*, Berlin Heidelberg, 2008, 882-891.

[16] F.J. Chang & Y.Y. Hwang, "A self-organization algorithm for real time flood forecast", *Hydrological Processes, 13*, 123-138.

[17] H.R. Madala and A.G. Ivakhnenko. "Inductive Learning Algorithms for Complex System Modeling", *Boca Raton*: CRC. 1974.

[18] N. Nariman-Zadeh, A. Darvizeh and F. Oliaei, "Genetic algorithm and singular value decomposition in the design of fuzzy systems for the modeling of explosive cutting process", *Proc. WSES Conf. Fuzzy Sets Fuzzy Systems, FSFS '02*, 2002.

[19] C.W. Hsu, C.C. Chang & C.J.Lin, "A Practical Guide for Support Vector Classification", http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.