Short-Term Electric Load Forecasting Using Standardized Load Profile (SLP) And Support Vector Regression (SVR)

Nguyen Tuan Dung
Planning Department,
EVNHCMC Power Company,
Ho Chi Minh, Vietnam
dp1526@gmail.com

Nguyen Thanh Phuong
Institute of Engineering,
Hutech University of Technology,
Ho Chi Minh, Vietnam
nt.phuong@hutech.edu.vn

Abstract—**Short-term load forecasting (STLF) plays an important role in business strategy building, ensuring reliability and safe operation for any electrical system. There are many different methods used for short-term forecasts including regression models, time series, neural networks, expert systems, fuzzy logic, machine learning, and statistical algorithms. The practical requirement is to minimize forecast errors, avoid wastages, prevent shortages, and limit risks in the electricity market. This paper proposes a method of STLF by constructing a standardized load profile (SLP) based on the past electrical load data, utilizing Support Regression Vector (SVR) machine learning algorithm to improve the accuracy of short-term forecasting algorithms.**

Keywords-short-term load forecast; regression model; standardized load profile; support vector regression

I. INTRODUCTION

Load forecasting is electrical systems is a topic that has been studied extensively. There are two main approaches in this area: Traditional statistical methods of the relationship between the load and load-affecting factors (such as time series, regression analysis, etc.) and machine learning methods (a branch of artificial intelligence). Statistical methods assume load data according to a sample and try to forecast the value of future loads using different time series analysis techniques. Intelligent systems are derived from mathematical expressions of human behavior and experience. Especially since the early 1990s, neural networks have been considered one of the most commonly used techniques in the field of electrical load forecasting, because they assume that there is a nonlinear function related to historical values and some external variables with future values may affect the output [1]. The approximate ability of neural networks has made their applications popular. In recent years, an intelligent calculation method involving Support Vector Machines (SVM) has been widely used in the field of load forecasting. Authors in [2] used the Support Vector Regression (SVR) technique to solve the electrical load prediction problem (forecasting a maximum daily load for the next 31 days). This was a competition organized by EUNITE (European Network on Intelligent Technologies for Smart Adaptive Systems). The provided information included: the demand data of the past two years, daily temperature of the past

four years and local holiday events. The data were divided into 2 parts: a part used for training (about 80%-90%) and the rest used for algorithm testing (about 20%-10%). The set of training inputs included data of previous day, previous hour, previous week, and the average of the previous week. Since then, there have been several studies exploring the different techniques used for optimizing SVR to perform load forecasting [3-10]. The main reason for using SVM in load forecasting is that it can easily model the load curve, the relationship between the load and the dynamics of changing load demand. However, there are some problems encountered when the above algorithms are applied to real situations:

- Climate conditions always play an important role in load forecasting. They show the relationship between climate and load demand. When we do load forecasting for the post-test period, it is very difficult to forecast the values of weather and climate used as the input of the algorithm and these values are often not available.

- Electrical load samples include hidden elements, which tend to be similar to the previous load model. However, it will lead to a false forecast of the following days if the date pattern is different from the previous day or there is an event that impacts. Therefore, the use of the dataset (training inputs include data of the previous day, the previous hour, the previous week, the average of the previous week) has many risks if the load models are not identical.

- If the forecast time frame is greater than the past data frame (more than 7 days), there will be a lack of input to run the algorithm.

- In addition, for Asian countries (such as Vietnam) that use lunar calendar, there are difficult and unpredictable issues as the Lunar New Year (usually in late January or early February), etc. There is a deviation between the solar calendar and the lunar calendar (the load models are not identical). Therefore, it often leads the forecast results of algorithm for this period with large errors.

This paper proposes a solution to build a Standardized Load Profile (SLP) based on the historical load dataset as a training

Corresponding author: Nguyen Tuan Dung

dataset. This input dataset is combined with the SVR algorithm to improve the accuracy of short-term forecast results, solve the problem of deviation between the solar and the lunar calendar, and overcome the input data frame. SLP will be built for all 365 days and 8,670 hour cycles in a year. SLP will be an important dataset during the training, testing and forecasting process. It will standardize load models by hours, days, seasons, and special day types (including lunar dates). Therefore, SLP will contribute to solve the above-mentioned difficulties and improve the quality of electrical load forecasting.

II. METHODOLOGY

Observing the load profiles of February of Ho Chi Minh City over the years (Figure 1), we can see a huge fluctuation in the chart shape over the years. The results in the use of historical data for forecasting this period of time are extremely complicated.
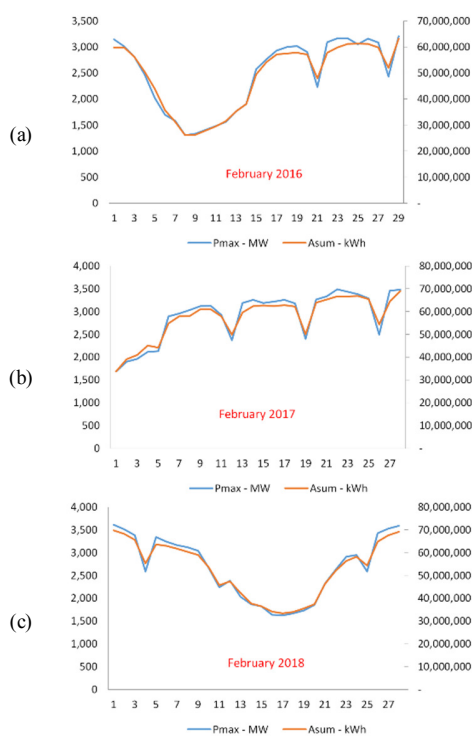

Fig. 1.   The load profiles of February over the years: (a) 2016, (b) 2017, (c) 2018

In fact, the algorithms used to forecast in Vietnam have to go through an intermediary stage in which the months are converted into regular months (without holidays and Lunar New Year). After, the forecast result will be reversed or the result will be accepted with a large error. This is a common problem in software provided by foreign countries.

A. Standardized Load Profiles (SLP)

While observing the load profiles of the days in a week and some special holidays of the year in Ho Chi Minh City (Figure 2), we see the difference between weekdays (from Tuesday to

Friday) is not much and they have the same load chart. For the load profiles on Monday, they are different from the normal days at 0:00 to 9:00, due to the forwarding demand from Sunday. For load profiles on Saturday, there is a change but not much compared to normal days. Mainly the load demand decreases in the evening due to the start of weekends. Particularly for load profiles on Sunday, it is completely different from normal days (the demand for electricity is low).
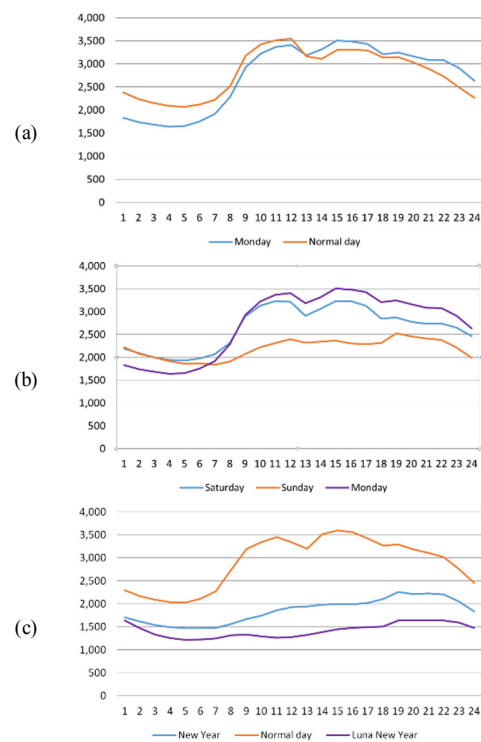

Fig. 2.   Typical load profiles on some days in a year

When observing the load chart of the New Year and the Lunar New Year, we see the difference completely, the graphs are almost flat, and the load demand is quite low because these are holidays. Particularly on Lunar New Year, the load demand is the lowest, because this is the longest holiday of the year (from 6 to 9 days). SLPs are built by taking the value of the collected capacity in a 60-minute period divided by its maximum capacity. We need to build SLP for 365 days per year. Some typical SLPs are shown in Figure 3. Based on the SLP of each cycle of the past data set, we can build the SLP data set for future forecast periods. This should be accurate to each cycle, each type of day (weekdays, working days, holidays, etc.), each week, and each month. Therefore, the SLP is a special feature and is also an important input parameter of the SVR (NN) training process to rebuild the load curves, from which we can estimate lost or not recorded data during the measurement process.

B. Support Vector Regression (SVR)

The feature of SVR is that it provides us with a sparse solution. That is, to build the regression function, we do not need to use all the data points in the training set. The points that

contribute to the construction of the regression are called support vectors. The layering for a new data point will depend only on the support vectors [5–6].
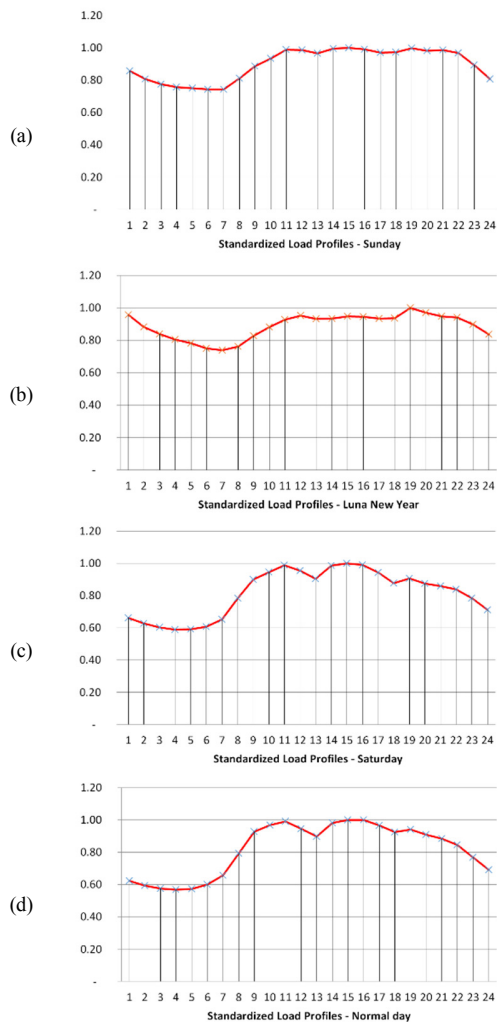

Fig. 3.   SLP of some days in a year: (a) Sunday, (b) Lunar New Year, (c) Saturday, and (d) a normal day.

The regression function has the formula:

$$y = f(x) = w^T \Phi(x) + b \qquad (1)$$

Thus, the goal of SVR training is to find w and b [7-10] for the training set $\{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\} \subset R^n \times R$. With a simple regression problem, to find w and b we have to minimize the normalized error function:

$$\frac{1}{2} \sum_{n=1}^{N} \{y_n - t_n\}^2 + \frac{\lambda}{2} \|w\|^2 \qquad (2)$$

where λ is a normalized constant.

To get a sparse solution, we will replace the above error function with the ε-insensitive error function. The characteristic of this error function is that if the absolute value of the

difference between the predicted value $y(x)$ and the target value is less than ε (with $\varepsilon > 0$) then the error is considered zero. Now, we must minimize the normalized error function:

$$C \sum_{n=1}^{N} E_\varepsilon (y(x_n) - t_n)^2 + \frac{1}{2} \|w\|^2 \qquad (3)$$

with $y(x_n) = f(x) = w^T \Phi(x_n) + b$, C is a normalized constant like λ but is multiplied by an error function instead of $\|w\|^2$.

To allow some points outside the tube ε, we will add slack variables. For each data point x_n, we need two liquid variables $\xi_v \geq 0$ and $\hat{\xi}_v \geq 0$, with $\xi_v \geq 0$ corresponding to the point that $t_n > y(x_n) + \varepsilon$ (outside and above the tube) and $\hat{\xi}_v \geq 0$ corresponding to the point that $t_n < y(x_n) - \varepsilon$ (outside and below the tube).
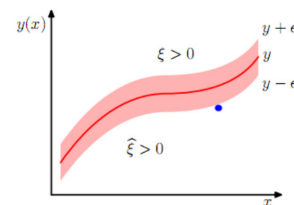

Fig. 4.   Illustration for liquid variables ξ_n

The condition for a destination point in the pipe is:

$y_n - \varepsilon \leq t_n \leq y_n + \varepsilon$ with $y_n = y(x_n)$. Using liquid variables, we allow destination points outside the tube (corresponding to liquid variables > 0) and thus the condition will now be:

$$t_n \leq y_n + \varepsilon + \xi_n$$
$$t_n \geq y_n - \varepsilon - \hat{\xi}_n$$

Thus, we have an error function for SVR:

$$C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n + \frac{1}{2} \|w\|^2)$$. Our goal is to minimize this

error function with constraints:

$$\xi_n \geq 0, \ \hat{\xi}_n \geq 0$$
$$t_n \leq y_n + \varepsilon + \xi_n$$
$$t_n \geq y_n - \varepsilon - \hat{\xi}_n$$

Using the Lagrange function and the Karush-Kuhn-Tucker condition, we have the equivalent optimization problem:

$$-\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \hat{a}_n)(a_m - \hat{a}_m) k(x_n, x_m)$$
$$-\varepsilon \sum_{n=1}^{N} (a_n - \hat{a}_n) + \sum_{n=1}^{N} (a_n - \hat{a}_n) t_n \qquad (4)$$

where k is the kernel function: $k(x, x') = \Phi(x)^T \Phi(x')$. maximizing with constraints:

$$0 \leq a_n \leq C$$
$$0 \leq \hat{a}_n \leq C \qquad\qquad (5)$$
$$\sum_{n=1}^{N} (a_n - \hat{a}_n) = 0$$

From here, we have the regression function of SVR:

$$y(x) = \sum_{n=1}^{N} (a_n - \hat{a}_n) k(x, x_n) + b \qquad (6)$$

Thus, for SVRs using the ε-insensitive error function and the Gaussian kernel function we obtain three parameters: the normalization coefficient C, the parameter γ of the Gaussian kernel function, and the width of the pipe ε [7]. These parameters affect the forecast accuracy of the model and need to be selected carefully.

- If C is too large, it will give a priority to the training error. It leads to a complex model and it is easy to be over fitting. If C is too small, it will give a priority to the complexity of the model. It leads to a too simple model and reduces forecast accuracy.

- The meaning of ε is the same. If it is too large, there will be less support vectors, making the model too simple. On the other hand, if ε is too small, there are many support vectors, leading to complex models, which are more likely to be over fitting.

- The γ parameter reflects the correlation between the support vectors and also affects the forecast accuracy of the model.

C. Research Models

The flowchart of the SLP-SVR forecasting algorithm is given in Figure 5.
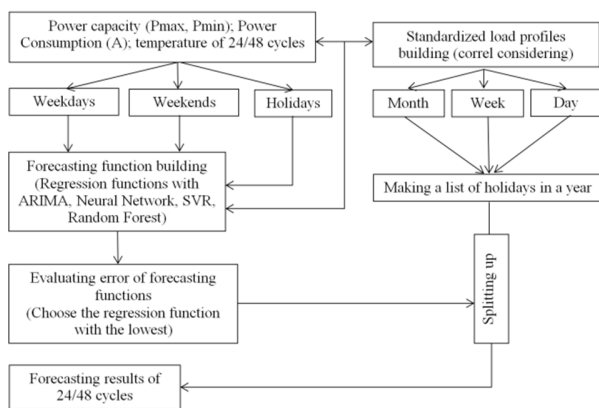

Fig. 5. Flowchart of forecasting algorithm by SLP – SVR

Processed historical data (power consumption, capacity and temperature recorded in 24 cycles - 60 minutes each) with the SLP will be included in modules to build regression functions under SVR, Neural Network (NN) algorithms to build

regression functions. Then we use the above dataset to check and evaluate the error of regression functions. After that, we choose the regression function with the smallest error to be used as regression function for the next forecast phase. The SLP dataset in 24 cycles of the expected period (including holidays, etc.) and the forecasted temperature in 24 cycles of the corresponding period will be the input for the regression function that is selected to export forecast results in 24 cycles for a period of 7-30 days.

III. RESULTS AND DISCUSSION

A. Input Data:

The article uses data from January 1st, 2015, to November 17th, 2018 of EVNHCMC to run test models. After pretreatment, the dataset is divided into 2 parts: training set and testing set, in which the testing set is the last 30 days of the dataset. Or the dataset is divided into phases to test the forecast results in different time periods. Input data for training algorithms include: capacity (Pmax/Pmin) in 60-minute cycles, temperature (max/min) in 60-minute cycles, standardized load profiles of 24 hours of day and a list of holidays and Lunar New Year in the forecast year. A useful measurement parameter is the mean absolute percentage error (MAPE) which is used to evaluate the error of models.

$$\text{MAPE} = \frac{1}{n} \sum 100 \frac{\left| Y_t - Y_t^f \right|}{Y_t} \qquad (7)$$

The algorithms are programmed in Matlab and the results are exported to Excel files for data exploitation.

B. SVR Models

It is necessary to correctly select the input parameters to run SVR models such as: normalization coefficient C, width of pipe ε and Gaussian kernel function. The algorithm uses the same input dataset of models. Some typical proposed SVR model parameters are shown in Table I.

TABLE I. SVR MODEL PARAMETERS

Model	C	ε	Kernel Function
SVR 1	93.42	32.5	Polynomial
SVR 2	500.32	0.01	Gaussian
SVR 3	1	50.03	Linear
SVR 4	100	0.01	Linear

C. RFR Models

A set of regression trees is used with each set of different rules to perform a non-linear regression. The algorithm builds a total of 20 trees, with a minimum leaf size of 20. The number of leaves is smaller or equal to the size of the tree to control overfitting and bring about high performance [13-14]. The algorithm uses the same input dataset of models.

D. Neural Network Models

We used Feedforward Neural Network models with the mentioned above input variables and training dataset. A-hidden-layer network architecture with a class size of 10 and Sigmoid activation function was used. At the same time, the usual Neural network with 3-hidden-layer network architecture,

in which: the first hidden layer has a size of 10 nodes, the second hidden layer has 8 and the third hidden layer has 5 nodes.

E. Results and Analysis

1) Regression Models Test

We run the forecast results for February of 2018 (the month of the Lunar New Year) to assess the degree of error of the models. The model included as inputs the data of the previous day, previous hour, previous week and the previous week average. Processed historical data (power consumption, capacity, temperature recorded at 24 cycles of 1 hour) with the SLP were included in modules to build regression functions under SVR, Neural Network and Random Forest algorithms to build regression functions.
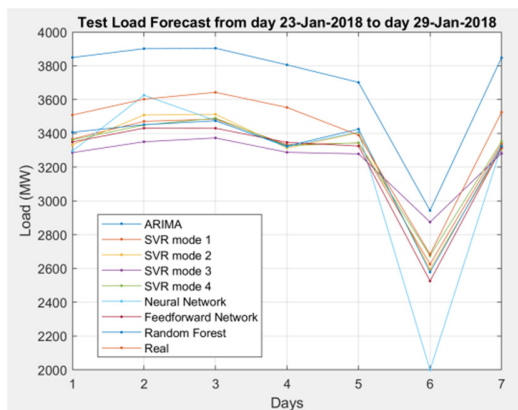

Fig. 6.   Regression models test

TABLE II.   CHECKING ERRORS OF REGRESSION MODELS RESULTS

Date	Ytr	Yts1	Yts2	Yts3	Yts4	YtNN	Ytfeed	YtRF
1/23/18	9.71	4.05	5.02	6.35	4.19	6.09	4.55	2.91
1/24/18	8.30	3.65	2.61	7.00	4.25	0.65	4.76	4.19
1/25/18	7.17	4.35	3.57	7.42	4.21	4.58	5.84	4.63
1/26/18	7.10	6.20	6.77	7.48	6.39	6.58	5.82	6.44
1/27/18	9.22	1.37	0.44	3.27	1.33	0.56	1.91	1.06
1/28/18	9.68	2.16	3.28	7.12	0.32	25.51	5.89	3.93
1/29/18	9.15	5.30	6.17	6.92	4.91	5.71	5.96	5.67

We chose the regression function with the smallest error to be used for the next forecast phase. The Yts4 model was selected as a forecasting model.

2) Forecast Results for February of 2018

Considering the model forecast results for February, we see a big difference between forecast and reality (Figure 7). The reason is that we used the historical data of January of 2019 (7-14-30 days before the forecasting date) as the input for the training model.

3) Results of Testing SVR Models

We see the results in Figure 8 and Table III.

4) Results of Testing Machine Learning Models
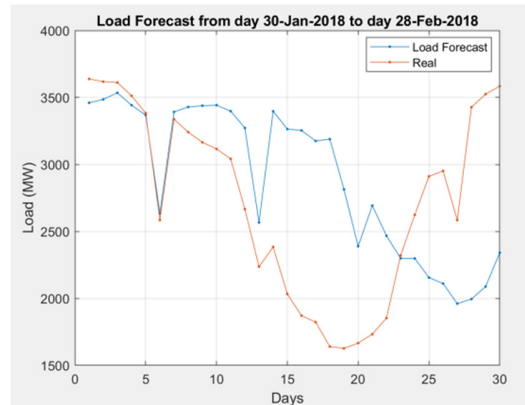
We see the results in Figure 9 and Table IV.

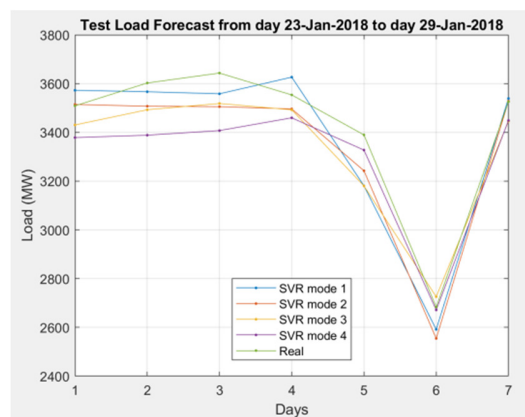Fig. 7.   Forecast results for the next 30 days

Fig. 8.   SVR models test

TABLE III.   RESULTS OF CHECKING ERRORS OF SVR MODELS

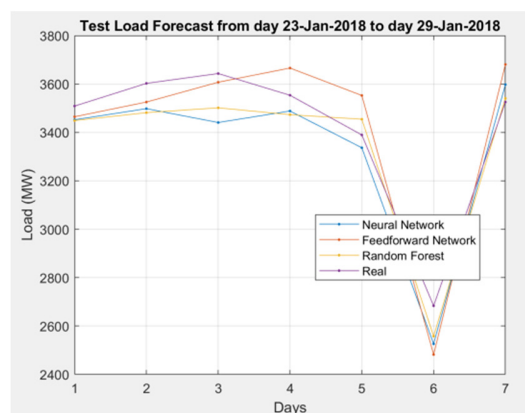| Date	Yts1	Yts2	Yts3	Yts4
1/23/18	1.15	0.64	2.22	3.87
1/24/18	1.70	2.12	2.95	6.19
1/25/18	3.03	3.30	3.38	6.68
1/26/18	1.35	1.04	1.76	2.76
1/27/18	6.77	4.56	6.42	1.56
1/28/18	4.18	5.09	1.81	0.76
1/29/18	0.24	0.12	2.69	2.14
MAPE	2.63	2.41	3.03	3.42

Fig. 9.   Machine learning models test

TABLE IV.   CHECKING ERRORS OF MACHINE LEARNING MODELS RESULT

Date	YtNN	YtFeed	YtRF
1/23/18	1.25	1.61	1.70
1/24/18	2.14	2.90	3.36
1/25/18	0.99	5.55	3.89
1/26/18	3.16	1.84	2.26
1/27/18	4.81	1.56	1.92
1/28/18	7.51	5.85	4.68
1/29/18	4.41	2.05	0.43
MAPE	3.47	3.05	2.60

5) Results of Testing Regression Models:

We see the results in Figure 10 and Table V.
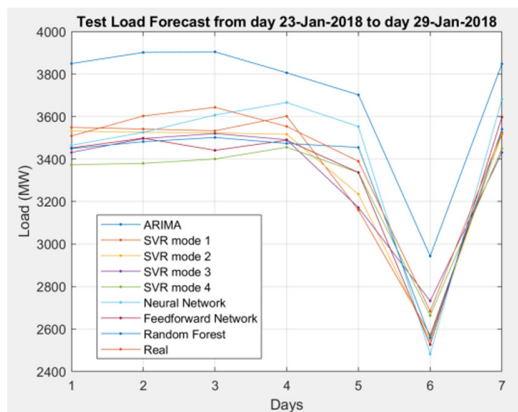

Fig. 10.   Regression test models

TABLE V.   RESULTS OF TEST MODELS CHECKING ERRORS

Date	Ytr	Yts1	Yts2	Yts3	Yts4	YtNN	Ytfeed	YtRF
1/23/18	9.71	1.15	0.64	2.22	3.87	1.25	1.61	1.70
1/24/18	8.30	1.70	2.12	2.95	6.19	2.14	2.90	3.36
1/25/18	7.17	3.03	3.30	3.38	6.68	0.99	5.55	3.89
1/26/18	7.10	1.35	1.04	1.76	2.76	3.16	1.84	2.26
1/27/18	9.22	6.77	4.56	6.42	1.56	4.81	1.56	1.92
1/28/18	9.68	4.18	5.09	1.81	0.76	7.51	5.85	4.68
1/29/18	9.15	0.24	0.12	2.69	2.14	4.41	2.05	0.43
MAPE	8.62	2.63	2.41	3.03	3.42	3.47	3.05	2.60

We choose the regression function with the smallest error to be used as the regression function for the next forecast phase. The model Yts2 is selected to be the forecasting model.

6) Forecast Results for February of 2018

We see the results in Figure 11, where a definite improvement is observed.

IV.   CONCLUSION

We observed the experimental results in the forms of testing datasets (load datasets of the previous day, previous week, previous month and the dataset of SLP), we saw that the results of the SLP-SVR models are closely to the actual value of February of 2018, while the results of the old model are in quite a large deviation. Thus, we see that the use of SLP as the input dataset for the modules of forecasting regression function is effective and gives forecasting results with low error. It solves the problem of deviation between the solar and the lunar dates, especially in the months of Lunar New Year. Also it resolves the difference between the solar and lunar cycles.
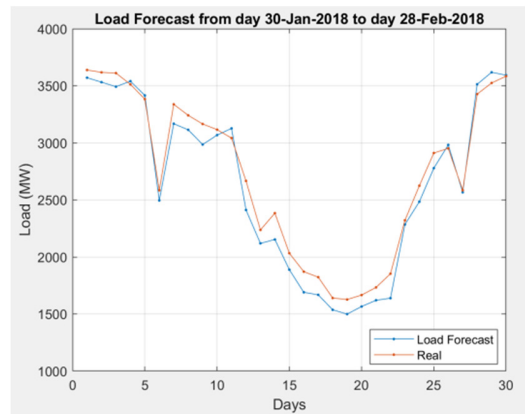

Fig. 11.   Forecast results for the next 30 days

REFERENCES

[1]   M. H. M. R. Shyamali Dilhani, C. Jeenanunt, "Daily electric load forecasting: Case of Thailand". 7th International Conference on Information Communication Technology for Embedded Systems, Bangkok, Thailand, March 20-22, 2016

[2]   J. Huo, T. Shi, J. Chang, "Comparison of Random Forest and SVM for Electrical Short-term Load Forecast with Different Data Sources", 7th IEEE International Conference on Software Engineering and Service Science, Beijing, China, March 23, 2017

[3]   L. C. P. Velasco, C. R. Villezas, P. N. C. Phalang, J. A. A. Dagaang, "Next Day Electric Load Forecasting   Using Artificial Neural Networks", Cebu City, Philippines, December 9-12, 2015

[4]   D. Willingham, "Electricity Load Forecasting for the Australian Market Case Study", available at https://ww2.mathworks.cn/matlabcentral/fileexchange/31877-electricity-load-forecasting-for-the-australian-market-case-study?s_tid=FX_rc1_behav, 2016

[5]   N. T. Dung, T. T. Ha, N. T. Phuong, "Comparative Study of Short-term Electric Load Forecasting: Case Study EVNHCMC", 4th International Conference on Green Technology and Sustainable Development, Ho Chi Minh City, Vietnam, November 23-24, 2018

[6]   E. Ceperic, V. Ceperic, A. Baric, "A strategy for short-term load forecasting by support vector regression machines", IEEE Transactions on Power Systems, Vol. 28, No. 4, pp. 4356-4364, 2013

[7]   V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995

[8]   S. Gunn, Support Vector Machines for Classification and Regression, Technical Report, University of Southampton, 1995

[9]   V. Cherkassky, Y. Ma, "Selection of Meta-parameters for Support Vector Regression", International Conference on Artificial Neural Networks, Madrid, Spain, August 28-30, 2002

[10]   D. Basak, S. Pal, D. C. Patranabis, "Support Vector Regression", Neural Information Processing – Letters and Reviews, Vol. 11, No. 10, pp. 203–224, 2007

[11]   A. J. Smola, B. Scholkopf, "A Tutorial on Support Vector Regression, Statistics and Computing", Vol. 14, No. 3, pp. 199–222, 2004

[12]   Understanding Support Vector Machine Regression and Support Vector Machine Regression, available at: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

[13]   L. Breiman, "Random Forests", Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001

[14]   L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees. Chapman & Hall 1984