The Comparison of PM2.5 forecasting methods in the form of multivariate and univariate time series based on Support Vector Machine and Genetic Algorithm

[1]Ronnachai Chuentawat
Informatics program, Faculty of science and technology
Nakhon Ratchasima Rajabhat University
Nakhon Ratchasima, Thailand
c_ronnachai@hotmail.com

[2]Yosporn Kan-ngan
Informatics program, Faculty of science and technology
Nakhon Ratchasima Rajabhat University
Nakhon Ratchasima, Thailand
k.yosporn@gmail.com

***Abstract*— This research aims to study and compare the forecasting precision of multivariate and univariate time series forecasting based on Support Vector Machine optimized with Genetic Algorithm. A study data is an hourly data set contains the PM2.5 data and a meteorological data in Beijing, China. This data is published in UCI Machine Learning Repository. This study starts with a generating of 3 data subsets from an original study data. Each data subset will be used to generate a multivariate and univariate time series model to forecast PM2.5. After that, we evaluate our research with error measurement by using RMSE and MAPE. From the results, we found that univariate time series models have lower error than multivariate time series models for all of 3 data subsets.**

Keywords—forecasting; multivariate time series; univariate time series; Support Vector Machine; Genetic Algorithm

I. INTRODUCTION

A time series data is recorded by time order and it has 2 types, a multivariate time series and a univariate time series. A multivariate time series has multiple observed variables, but a univariate time series has a single observed variable. A time series data are deployed to study and research in many fields and one of the fields is a forecasting time series. To generate a forecasting model, there are many different techniques, if we use a suitable technique with a characteristic of time series, we will get a precisely forecasting model. An accurately forecasting will bring through many advantages such as good performance of production planning, decreasing operating cost or planning to protect future problem efficiently. This research focus on a study to generate a multivariate and a univariate forecasting time series model. The data set is used in this study came from UCI Machine Learning Repository, it's called that Beijing PM2.5 Data Set. This data is a multivariate time series, it's an hourly data set contains the concentration values of small particle matter less than 2.5 microns in the air (PM2.5) and meteorological data in Beijing, China.

Techniques to generate a forecasting model may be classified into 2 groups, the first group is the traditional statistic techniques [1], such as regression analysis, exponential smoothing or Box and Jenkins method [2] to generate the ARIMA model. Most of techniques in the first group are always used to forecast a univariate time series, but may except for a regression analysis, which can be done as a multiple regression [3]. The second group is the machine learning techniques [4] such as an Artificial Neural Network (ANN) or Support Vector Machine (SVM). The machine learning techniques can be used to generate both of multivariate and univariate time series forecasting model. When we use them to generate a multivariate model, inputs of a model are variables or factors relate to a forecasting variable as an output, but if we use them to generate a univariate model, inputs of a model are lag time observed values and the output is an observed value at time t.

Support Vector Machine (SVM) is one of the machine learning techniques and it was introduced by Vapnik [5]. The SVM is a popular technique to be used in data mining or machine learning fields because it has a problem of overfitting less than other techniques such as the artificial neural network technique [6,7]. The SVM is always used to classify data by generating the optimal hyperplane to classify data into 2 classes. When the SVM is used to forecast data, it's called that Support Vector Regression (SVR). The SVR will generate the epsilon tube around the optimal hyperplane and this tube is used to be a relative representation of an input vector in n dimension ($X \in Rn$) and one output variable ($y \in R$). The forecasting equation of SVR can be shown as equation 1 [8].

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

Where $\alpha_i, \alpha_i^* > 0$ are the positive Lagrange multipliers, $k(x_i, x)$ is a kernel function used to map data into high dimensional space and b is a constant vector and is called as a bias. The SVM has some parameters and they should be determined with optimal value. To determine optimal value, we want an algorithm to search it and one of the searching algorithm is the Genetic Algorithm (GA), it was introduced by Holland [9]. The GA is a searching algorithm that uses a concept of evolution theory. The GA starts with initial population and member of the population is called that chromosome. Each chromosome contains some genes and each gene is an optimal value of each parameter. Each chromosome is evaluated by a fitness function and is operated by genetic operations contain with a selection, a crossover and a mutation operation respectively. After genetic operation, we will get a new population is called offspring that

has good fitness more than its parent (previous population). The step of the GA will be repeated until last generation and return the optimal parameter finally. Therefore, this research uses the SVM to generate PM2.5 forecasting models in the form of multivariate and univariate time series and uses the GA to determine the optimal parameters of the SVM. Finally, we evaluate our experiment by accuracy measurement between a multivariate and a univariate of the PM2.5 forecasting model.

II. METHODOLOGY

This research aims to compare the accuracy of multivariate and univariate of the PM2.5 forecasting model, when the forecasting models are generated by the SVM and optimized its parameters by GA. We used the Beijing PM2.5 Data Set that is used in a prior research of Liang et. al. [10]. This data is a multivariate time series and has hourly time periods during Jan 1st, 2010 to Dec 31th, 2014 with a number of instances equal to 43,824. When we use this data to generate a multivariate forecasting model based on SVM which is a technique that use with only numerical data, therefore, we select only numerical input variables to forecast the PM2.5 output that consist of a dew point (DEWP), a temperature (TEMP), a pressure (PRES) and a cumulated wind speed (Iws). Therefore, our multivariate time series model can be shown as Fig. 1. When we use this data to generate a univariate forecasting model, we observe values of PM2.5 only and inputs of the model are lag time values of PM2.5 while an output is a value at time t. We present the method to select inputs of our univariate time series model by using the coefficient of determination (R^2) between 1 to 12 lag time values and value of PM2.5 at time t, after that, we choose the top 4 of R^2 to be 4 inputs of the model. Therefore, our univariate time series model can be shown as Fig. 2.

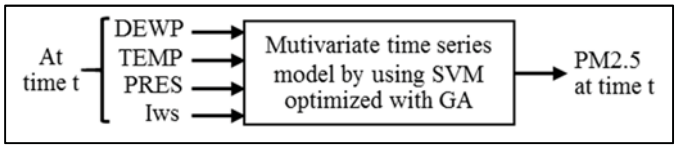

Fig. 1. Multivariate time series model

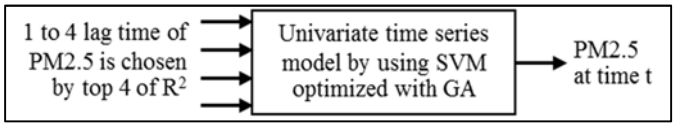

Fig. 2. Univariate time series model

We start our experiment by generating 3 data subsets from the Beijing PM2.5 Data Set. Three data subsets were chosen for least missing values and eliminate them by finding the average value between previous and next value. The first subset is a data during Dec 29th, 2012 to May 16th, 2013 and has a number of instances equal to 3,336. The second subset is a data during Oct 23th, 2013 to Dec 28th, 2013 and has a number of instances equal to 1,608. The third subset is a data during Mar 1st, 2014 to Apr 7th, 2014 and has a number of instances equal to 912. Each data subset is experimented with the same procedure and experimental procedure can be shown as Fig. 3. From Fig. 3, we can summarize our experiment as follows, this experiment generates a multivariate and univariate time series model based on SVM and GA, which has an algorithm to find optimal solution as shown in Fig. 4, to forecast PM2.5 and evaluates experiment by error measurement. We measure the error of a multivariate and univariate model by using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The RMSE and MAPE can be calculated as equation 2 and 3 respectively [11]

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (2)$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left| 100\,\frac{y_t - \hat{y}_t}{y_t}\right| \quad (3)$$

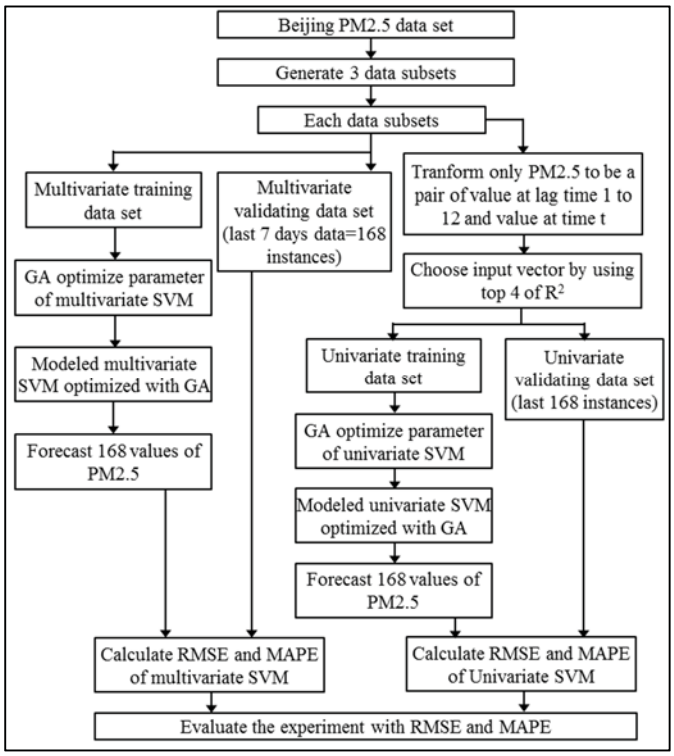

Fig. 3. Experimental procedure

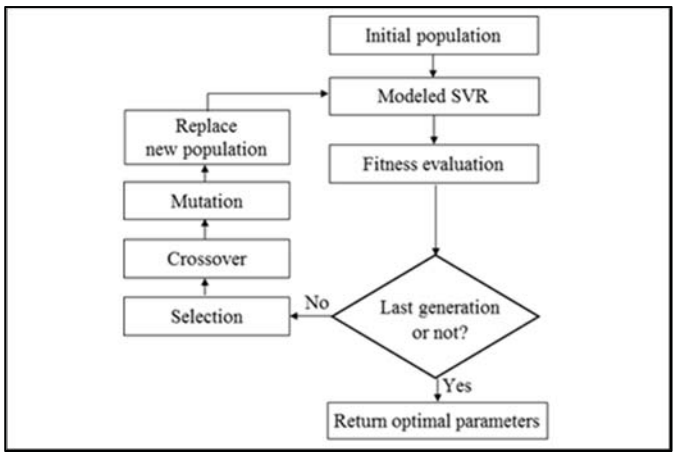

Fig. 4. Flow chart of Genetic Algorithm

III. Experimental Results

This research uses R language as an experimental tool and initial experiment by generating 3 data subsets from the Beijing PM2.5 Data Set. Each data subset will be used to perform an experiment with a same procedure and starts an experiment by the coefficient of determination (R^2) analysis between inputs and output of a multivariate and univariate time series model. In R language, we can use the lm() function, which is a function to generate a simple linear regression model, to summarize R^2 between a pair of inputs or independent variables and output or dependent variable. Our experimental results found that the inputs of univariate time series model that have the top 4 of R^2 values are lag time values of PM2.5 at lag time t-1, t-2, t-3, and t-4 respectively and 3 data subsets have a same result. For all of multivariate time series models, there are 4 clearly inputs, including a dew point (DEWP), a temperature (TEMP), a pressure (PRES) and a cumulated wind speed (Iws). The R^2 of multivariate and univariate time series model for all of 3 data subsets can be shown as table I.

TABLE I. THE COEFFICIENT OF DETERMINATION VALUES

Data subsets	Multivariate time series				Univariate time series			
	DEWP	TEMP	PRES	Iws	t-1	t-2	t-3	t-4
1	0.12	0.03	0.01	0.08	0.94	0.86	0.79	0.72
2	0.39	0.00	0.07	0.10	0.94	0.84	0.74	0.65
3	0.48	0.00	0.18	0.09	0.95	0.87	0.80	0.73

When we clearly determine the inputs and output of a forecasting model, it's mean that we know the structure of a model and can generate a model by dividing the original data subsets into a training set to train a model and validating set, which is a set of PM2.5 in the last 7 days, to validate the accuracy of a model. A training set will be used to train a model by the SVM technique and the optimal parameters of the SVM will be determined by the GA. In R language, there is the rgba() function to search optimal solution follow by an algorithm of the GA. From random experiment, we found that the best kernel of the SVM is a linear kernel that has two parameters, including a cost (C) and epsilon (ε). When we use the GA to search the optimal parameters of the SVM, each SVM has its optimal parameters as shown in table II and a sample result of the GA for the first data subset of a multivariate time series can be shown as Fig. 5. When we know the optimal parameters, we can generate the optimal SVM by using the svm() function in R. After we generate the SVM model already, we use it to forecast PM2.5 for last 7 days of each data subset and compare them with the actual values at the same time by calculating RMSE and MAPE. The RMSE and MAPE of each model can be shown as table III.

TABLE II. THE OPTIMAL PARAMETER OF THE SVM

Data subsets	Multivariate time series		Univariate time series	
	C	ε	C	ε
1	6.0820	0.1177	1.7482	0.2616
2	12.5394	0.1614	15.8225	0.9348
3	24.3147	0.1130	27.4656	0.5051

TABLE III. THE RMSE AND MAPE

Data subsets	Multivariate time series		Univariate time series	
	RMSE	MAPE	RMSE	MAPE
1	56.18	98.33	27.03	35.54
2	126.54	91.45	38.48	21.71
3	57.94	134.57	13.71	22.85

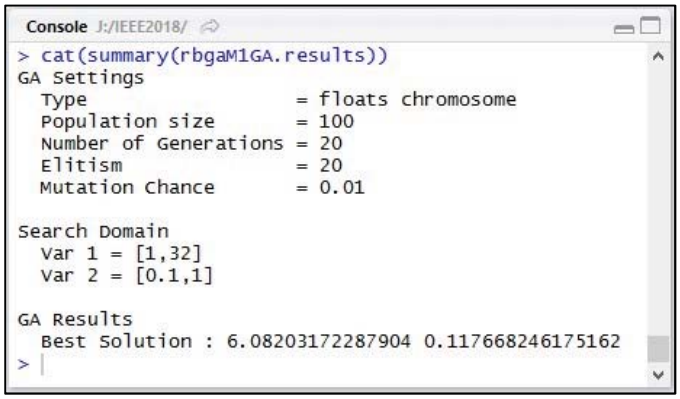

Fig. 5. Result of GA for the first multivariate time series

Finally, when we plot forecasting values versus the actual values for all of the forecasting models, the result can be shown as Fig. 6 to 8.

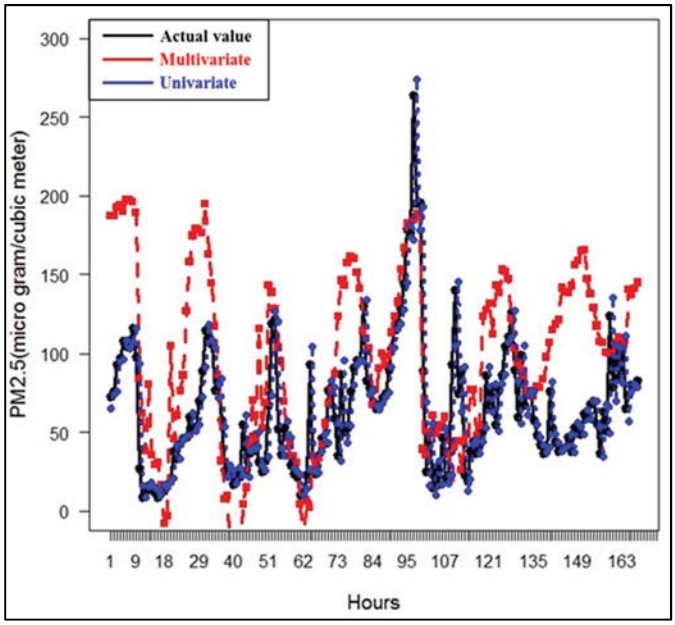

Fig. 6. Comparing graph of the first data subset

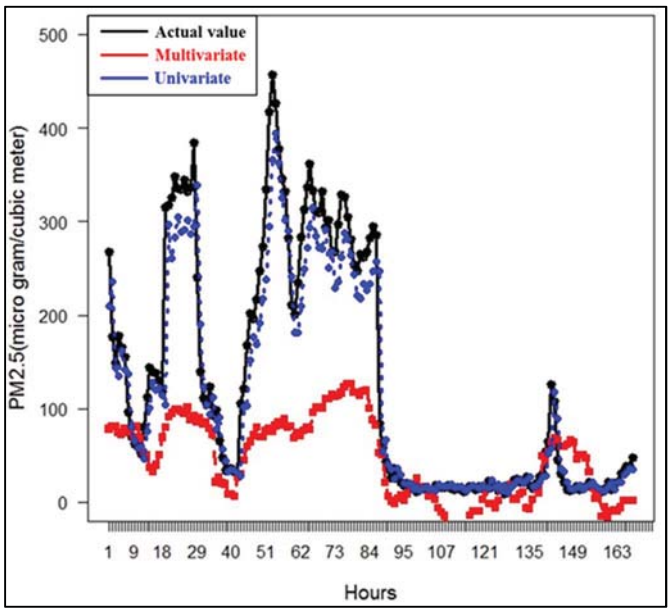

Fig. 7. Comparing graph of the second data subset

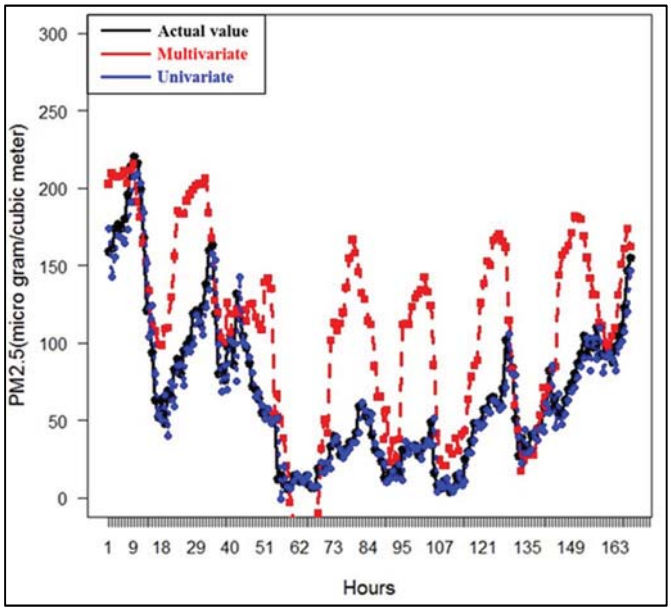

Fig. 8. Comparing graph of the third data subset

IV. Conclusion

This research aims to study and compare the PM2.5 forecasting methods in the form of multivariate and univariate time series based on Support Vector Machine (SVM) and Genetic Algorithm (GA). This research starts by generating 3 data subsets from the Beijing PM2.5 Data Set that was published in UCI Machine Learning Repository. Each data subset is used to generate a multivariate and univariate forecasting model. The inputs for all of multivariate forecasting models are other numeric variables in the Beijing PM2.5 Data Set, consisting of a dew point (DEWP), a temperature (TEMP), a pressure (PRES) and a cumulated wind speed (Iws). The inputs for all of univariate forecasting models are determined by R^2 analysis, including 4 lag time of the PM2.5 value at lag time t-1, t-2, t-3 and t-4 respectively. When we evaluate the accuracy of our multivariate and univariate forecasting models by measuring Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), we found that a univariate forecasting model of each data subset has lower RMSE and MAPE than a multivariate forecasting model. When we analyze the R^2 between inputs and output, we found that all R^2 values of univariate models have very greater than all R^2 values of multivariate models. When we plot the forecasting values of univariate models and the forecasting values of multivariate models, compare with the actual values, we found that a graph of univariate forecasting values very similar with a graph of actual values, but a graph of multivariate forecasting values non similar with a graph of actual values for all of 3 data subsets. From these results, we can conclude that the relation between inputs and output has more affect to the model accuracy than number of variable to generate the model and the univariate time series modeling that determine inputs to relate with output has more accurate and detect a movement of data better than the multivariate time series modeling that has inputs not relate to output.

References

[1] V. Prema, and K.U. Rao, "Development of statistical time series models for solar power prediction," Renewable Energy, vol. 83, pp. 100-109, 2015.

[2] G.E.P. Box, and G. Jenkins, Time Series Analysis, Forecasting and Control, 3rd ed. San Francissco: Holden-Day, 1990.

[3] K. Kerdprasop, and N. Kerdprasop, "Rainfall Estimation Models Induced from Ground Station and Satellite Data," Proceedings of the International MultiConference of Engineers and Computer Scientists, March 2016.

[4] S.C. James, Y. Zhang, and F. O'Donncha, "A machine learning framework to forecast wave conditions," Coastal Engineering, vol. 137, pp. 1-10, 2018.

[5] V. Vapnik, The Nature of Statistical Learning Theory, 1st ed. Springer-Verlag, 1995.

[6] G. Ogcu, O.F. Demirel, and S. Zaim, "Forecasting electricity consumption with neural networks and support vector regression," Procedia-Social and Behavioral Sciences, vol. 58, pp. 1576-1585, 2012.

[7] P.F. Pai, and C.S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," Omega, vol. 33, pp. 497-505, 2005.

[8] P. Bagheripour, A. Gholami, M. Asoodeh, and M.V. Asadi, "Support vector regression based determination of shear wave velocity," Journal of Petroleum Science and Engineering, vol. 125, pp. 95-99, 2015.

[9] J.H. Holland, Adaptation in Natural and Artificial Systems, 1st ed. Michigan: University of Michigan Press, 1975.

[10] X. Liang, T. Zou, B. Guo, S. Li, H. Zhang, S. Zhang, H. Huang, and S.X. Chen, "Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating," Proceedings of the Royal Society A, October 2015.

[11] C. Bergmeir, and J.M. Benítez, "On the use of cross-validation for time series predictor evaluation," Information Sciences, vol. 191, pp. 192-213, 2012.