TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 9, September 2014, pp. 6946 ~ 6953
DOI: 10.11591/telkomnika.v12i9.4639

Short-term Traffic Flow Prediction Based on Multivariable Phase Space Reconstruction and LSSVM

Duo Zhang*, Fengqing Han
School of Management, Chongqing Jiaotong University, Chongqing, 400074, China
*Corresponding author, e-mail: cqzhangd2012@163.com

Abstract

Real-time and accurate short-term traffic flow prediction is the premise and key of intelligent traffic control and guidance system. According to this problem, this paper put forward a prediction model based on multivariable phase space reconstruction and least squares support vector machine (LSSVM). First, the model confirms embedding dimension and delay time of the traffic flow, occupancy and average speed time series by analyzing their chaotic characteristics, and reconstructs multivariable state space. Second, the phase points obtained after reconstruction are as input, and the last traffic flow parameters came from the following phase points are as output. Finally, the LSSVM which be trained is adapted to realize short-term traffic flow prediction. This research compares this model with a model based on univariate phase space reconstruction and LSSVM, and the results show that the model proposed in this paper predicts better.

Keywords: *multivariable, chaos, phase space reconstruction, traffic flow prediction*

1. Introduction

Intelligent transportation system is considered as an effective method to handle transportation issues, such as alleviate traffic congestion, and reduce traffic accidents and vehicle emissions. Traffic control and guidance system is an integral part of intelligent transportation system, and how to quickly and accurately forecast short-term traffic flow is the emphasis and difficulty of traffic control and guidance system development. Currently, there are mainly three kinds of short-term traffic flow prediction methods: 1) Based on statistical theory: such as the historical average, Kalman filtering method and non-parametric regression method, etc.; 2) Based on the theory of nonlinear prediction: such as method based wavelet theory, fractal theory, cusp catastrophe theory, etc.; 3) Based on the theory of intelligent methods: such as neural networks and support vector machine [1- 4].

Intelligent transportation system is a complex, time-varying and high-dimensional nonlinear system, and its distinguishing feature is high degree of uncertainty. For these features of transportation system, support vector machine is used to forecast short-term traffic flow by some experts and scholars [2-4]. Because of Support Vector Machine (SVM) is a machine learning method based on the structural risk minimization and statistical learning theory, it can effectively solve small sample, nonlinearity, high dimension and local minima problems. In addition, some studies show that traffic flow has chaotic characteristics; chaotic exists in traffic flow by sampling data analysis [5]. Chaotic exists in traffic flow of 15 minutes and 5 minutes [6]. As a result, more and more researcher applies support vector machine and chaotic time series theory to study short-term traffic flow prediction problem [7, 8]. LSSVM and phase space reconstruction theory are adopted to predict urban traffic flow, and the data which be used is one-dimensional traffic flow time series [7]. Although Euclidean distance and similarity are taken into consideration to select the point nearby, the main idea is predicting the reconstructed time series multi-step by the LSSVM [8]. In summary, it is clearly that only one variable time series is used in the researches above. However, the actual transportation system is described by a number of variables, so multivariate time series model may contain more dynamic information, and especially when there are some noises in the system, the model can filter out noise and improve the prediction accuracy.

This paper proposes a prediction model based on Multivariable Phase Space Reconstruction (MPSR) and LSSVM. Specifically, the model confirms embedding dimension and delay time of the traffic flow, occupancy and average speed time series by analyzing their

Received November 6, 2013; Revised June 1, 2014; Accepted June 17, 2014

chaotic characteristics, and then reconstructs multivariable state space and fits the reconstructed time series by using LSSVM. As a result, this paper realizes short-term traffic flow prediction.

2. Multivariate Phase Space Reconstruction

2.1. Basic Principles of Multivariate Phase Space Reconstruction

Suppose M-dimensional multivariate time series $X_1, X_2, X_3, \ldots, X_{M-1}, X_M$, where $X_i = \left(x_{i,1}, x_{i,2}, \ldots, x_{i,N}\right),\ i = 1, 2, \ldots, M$, and reconstruct the multivariate time series.

$$\begin{aligned} D(n) = \Big(\ & x_{1,n}, x_{1,(n+{}^{2}{}_{1})}, \ldots, x_{1,(n+(d_1-1){}^{2}{}_{1})} \\ & x_{2,n}, x_{2,(n+{}^{2}{}_{2})}, \ldots, x_{2,(n+(d_2-1){}^{2}{}_{2})} \\ & \qquad \vdots \\ & x_{M,n}, x_{M,(n+{}^{2}{}_{M})}, \ldots, x_{M,(n+(d_M-1){}^{2}{}_{M})} \Big) \end{aligned} \tag{1}$$

Where $n = 1, 2, \cdots, \min_{1\le i\le M} N - \left(d_i - 1\right){}^{2}{}_{i}$, ${}^{2}{}_{i}$ is delay time, $d_i \left(i = 1, \ldots, M\right)$ is embedding dimension.

According to Tankens' delay embedding theorem [9], if d or d_i is sufficiently large, there is a mapping F: $R^d \to R^d$, where $d = \sum_{i=1}^{M} d_i$, and F is smooth function on d-dimensional space, which F can make:

$$D(n+1) = F\left(D(n)\right) \tag{2}$$

The above equation can also be written as:

$$x_{1,n+(d_1-1){}^{2}{}_{1}+1} = F_1\left(D(n)\right) \tag{3}$$

2.2. Confirmation of Embedding Dimension and Delay Time

Seen from the above, delay time ${}^{2}{}_{i}$ and embedding dimension d_i are the two parameters of phase space reconstruction. Parameters ${}^{2}{}_{i}$ and d_i are determined separately in conventional methods. In recent years, some studies show that the main factors affecting the quality of the phase space are not how to select ${}^{2}{}_{i}$ and d_i individual, but it is more important that the factors are determined by the embedding window width $h_i = \left(d_i - 1\right){}^{2}{}_{i}$. In 1999, Kim, Eykholt and Salas proposed C-C method [10], this method can estimate the delay time ${}^{2}{}_{i}$ and embedding window width h_i through correlation integral simultaneously.

$$\begin{cases} S_{cor}\left(t_i\right) = \Delta\bar{S}(t) + \left|\bar{S}(t)\right| \\ \Delta\bar{S}(t) = \dfrac{1}{4}\displaystyle\sum_{m=2}^{5} \Delta S\left(m, N, t\right) \\ \bar{S}(t) = \dfrac{1}{16}\displaystyle\sum_{m=2}^{5}\sum_{j=2}^{4} S\left(m, N, r_j, t\right) \end{cases} \tag{4}$$

Where $\Delta S(m, N, t) = \max\left[S\left(m, N, r_i, t\right)\right] - \min\left[S\left(m, N, r_j, t\right)\right] \quad i \ne j \quad S\left(m, N, r, {}^{2}\right)$ is the statistic of each sub-sequence, $S\left(m, N, r, {}^{2}\right) = \dfrac{1}{t}\sum_{l=1}^{t}\left\{C_l\left(m, N/r, r, {}^{2}\right) - \left[C_l\left(1, N/t, r, {}^{2}\right)\right]^m\right\}$; $C\left(m, N, r, {}^{2}\right)$ is the correlation integral; The first minimum value of $\Delta\overline{S}(t)$ is the delay time of the time series, and the minimum value of $S_{cor}(t)$ is the embedding window width of the time series.

3. Prediction Model of LSSVM

Least squares support vector regression machine is the application of LSSVM in regression estimation proposed by Suykens J.A.K in 1999 [11]. With quadratic loss function, LSSVM switch inequality constraints in the traditionally support vector machines into equality constraints. What's more, it takes squared error and loss function as loss experience of the training set. Therefore, the optimization problem is transformed into a linear equation problem. Since it is used to solve linear equation problem, this method presents the advantages of less computational resources, faster solution and faster convergence when the data is large.

3.1. Basic Principles of LSSVM

Suppose nonlinear sample data:

$$(x_1,y_1),(x_2,y_2),...,(x_i,y_i),...(x_l,y_l) \tag{5}$$

Where $x_i \in R^n, y_i \in R, i = 1,2,...,l$.

Set nonlinear function:

$$f(x) = \text{µ} \cdot \text{¨}(x) + b \tag{6}$$

According to statistical learning theory, function estimation problem can be described as the following optimization problem.

$$\begin{aligned} &\min \quad \frac{1}{2}\|\text{µ}\|^2 + \text{¥}\sum_{i=1}^{l}\text{¶}_i^2 \\ &s.t. \quad y_i = \text{µ}^T\text{¨}(x_i) + b + \text{¶}_i, i = 1,2,...,l \end{aligned} \tag{7}$$

Where $\text{¨}(x): R^n \to H$ is the mapping function of feature space, µ is the weight vector of the space H , $b \in R$ is the deviation, ¶_i is the error variable, and ¥ is the real-valued constant.

The above formula is a convex quadratic optimization problem. Introducing Lagrange function:

$$L(\text{µ},b,\text{¶},a) = \frac{1}{2}\|\text{µ}\|^2 + \text{¥}\sum_{i=1}^{l}\text{¶}_i^2 - \sum_{i=1}^{l}a_i\left[\text{µ}^T\text{¨}(x_i) + b + \text{¶}_i - y_i\right] \tag{8}$$

Where $a_i \geq 0$ is the Lagrange multiplier.

According to the KKT conditions, there is:

$$\begin{cases} \dfrac{\partial L}{\partial \text{µ}} = 0 \Rightarrow \text{µ} = \sum_{i=1}^{l} a_i \text{¨}(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l} a_i = 0 \\ \dfrac{\partial L}{\partial \text{¶}_i} = 0 \Rightarrow a_i = \text{¥}e_i \\ \dfrac{\partial L}{\partial a_i} = 0 \Rightarrow \text{µ}\text{¨}(x_i) + b + \text{¶}_i - y_i = 0 \end{cases} \tag{9}$$

Use formula (9) to eliminate µ and $¶_i$, and get the implementation form of the LSSVM.

$$\begin{bmatrix} 0 & 1_{1\times l} \\ 1 & ¨\left(x_i\right)^T ¨\left(x_j\right)+\frac{1}{¥}I \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{10}$$

Where $1_{1\times l}=[1,1,...,1]$, $a=[a_1;a_2;...;a_l]$, $i,j=1,2,...,l$, I is the unit matrix, a_i and b can be solved by the formula (10).

Therefore, nonlinear decision function can be obtained, where $k\left(x\cdot x_i\right)=¨\left(x\right)^T ¨\left(x_i\right)$.

$$f\left(x\right)=\sum_{i=1}^{l}a_i k\left(x\cdot x_i\right)+b \tag{11}$$

3.2. Selection and Optimization of Parameters

The performance of LSSVM depends on the selection of parameters, and the specific model is uniquely determined by its parameters. Because this paper uses RBF kernel function, LSSVM just determine the kernel function’s parameter and the penalty factor, and no longer need to determine insensitive loss coefficient, As a result, it simplifies the model structure.

At present, the selection of these two parameters is based on experience method and cut and tries method. This paper uses particle swarm optimization algorithm to optimize parameters of LSSVM to find the best parameter combination.

3.3. Prediction Procedure of Traffic Flow

Step 1: Use C-C method mentioned in literature [10] to determine the embedding dimension d_1,d_2,d_3 and delay time $²_1,²_2,²_3$ of the traffic flow, occupancy and average speed time series $X_i=\left(x_{i,1},x_{i,2},\ldots,x_{i,N}\right),i=1,2,3$.

Step 2: Calculate the average period p_1,p_2,p_3 of the time series X_1,X_2,X_3 by Fast Fourier Transform. Then use the largest Lyapunov exponent method [12] to judge the chaotic characteristics of traffic flow, occupancy and average speed time series

Step 3: Get the phase points by reconstructing multivariable state space of traffic flow, occupancy and average speed time series.

$$\begin{aligned} D(n)=\Big(& x_{1,n},x_{1,(n+²_1)},\ldots,x_{1,(n+(d_1-1)²_1)} \\ & x_{2,n},x_{2,(n+²_2)},\ldots,x_{2,(n+(d_2-1)²_2)} \\ & x_{3,n},x_{3,(n+²_3)},\ldots,x_{3,(n+(d_3-1)²_3)}\Big) \end{aligned} \tag{12}$$

Where $n=1,2,\cdots,L$, $L=\min_{1\le i\le 3}N-\left(d_i-1\right)²_i$.

Step 4: After the phase space reconstruction, the phase point obtained after reconstruction is inputted, and the last traffic flow parameter came from the following phase point is outputted. Then normalize the inputs and outputs and divide them into training set and test validation set. The inputs and outputs of prediction model based on phase space and LSSVM is expressed as follows:

$$\text{Inputs}=\begin{cases} D_1=\left(x_{1,1},\cdots,x_{1,1+(d_1-1)²_1},\cdots,x_{3,1},\cdots,x_{3,1+(d_3-1)²_3}\right) \\ D_2=\left(x_{1,2},\cdots,x_{1,2+(d_1-1)²_1},\cdots,x_{3,2},\cdots,x_{3,2+(d_3-1)²_3}\right) \\ \qquad\vdots \\ D_{L_1}=\left(x_{1,L_1},\cdots,x_{1,L_1+(d_1-1)²_1},\cdots,x_{3,L_1},\cdots,x_{3,L_1+(d_3-1)²_3}\right) \end{cases} \tag{13}$$

$$\text{Outputs} = \begin{cases} Y_1 = x_{1,1+(d_1-1)^{2}{}_1} \\ Y_2 = x_{1,2+(d_1-1)^{2}{}_1} \\ \vdots \\ Y_{L_1} = x_{1,L_1+(d_1-1)^{2}{}_1} \end{cases} \tag{14}$$

Step 5: Train LSSVM, and get traffic flow forecasting model. Then use the trained LSSVM to predict the test validation set.

4. Case Study

To verify the feasibility and the prediction accuracy of the model, this study adopts five minutes aggregate data of the sensor NO.601208 of California PeMS 12.3 system for model validation. The data includes three parameters -traffic flow, occupancy, average speed. The paper select 864 sets of data from June 17, 2013 to June 17 2013 as the training set, and 288 sets of data from June 20, 2013 as the test validation set. Completing this example verification by Matlab7.6 on PC with CPU 2.53GHZ, memory 2G, and Windows7 operating system.

4.1. Confirmation of Embedding Dimension and Delay Time

This paper adopts the C-C method to determine the embedding dimension and delay time of the traffic flow, occupancy and average speed time series. The calculation results are shown in Figure 1 to Figure 3. The first minimum value of $\Delta\bar{S}(t)$ is the delay time of the time series, and the minimum value of $S_{cor}(t)$ is the embedding window width of the time series. Therefore, according to Figure 1 to Figure 3, the Table 1 can be obtained.

Table 1. Embedding Dimension and Delay Time

	traffic flow	occupancy	average speed
embedding dimension	$d_1 = 7$	$d_2 = 6$	$d_3 = 9$
Delay time	$^{2}{}_1 = 18$	$^{2}{}_2 = 22$	$^{2}{}_3 = 15$

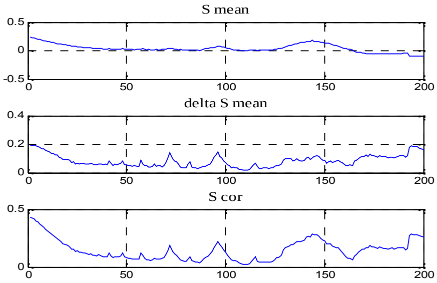

Figure 1. C-C Algorithm Results-traffic Flow

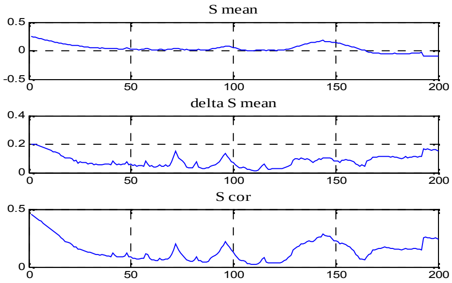

Figure 2. C-C Algorithm Results- occupancy

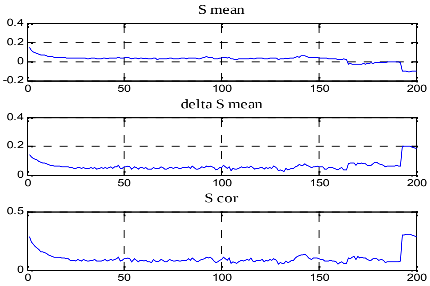

Figure 3. C-C Algorithm Results- average Speed

4.2. Calculation of Largest Lyapunov Exponent

According to the above results and the average periods $p_1 = p_2 = p_3 = 288$ obtained by Fast Fourier Transform, we adopt the method of small data sets to analyze traffic flow, occupancy and average speed time series, and the calculation results are shown in Figure 4. The straight lines of the Figure 4 are the fitting straight lines of the least squares method, whose slopes are the largest Lyapunov exponents. The largest Lyapunov exponents of traffic flow, occupancy and average speed time series are 0.0011, 0.0014 and 0.0012, respectively. Because all of the largest Lyapunov exponents are greater than zero, we get the conclusion that traffic flow, occupancy and average speed time series show the chaotic characteristics.

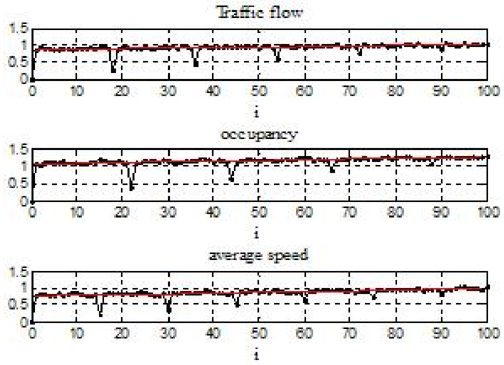

Figure 4. The Largest Lyapunov Exponent

4.3. Prediction of Short-term Traffic Flow

In order to compare the MPSR-LSSVM model (Prediction Model of LSSVM based on MPSR) with ASV-LSSVM model (Prediction Model of LSSVM Based on A Singe Variable) mentioned in literature [7], this paper draws comparative diagrams of predicted values and actual values (only show 100 sets data for clear comparison). The results are shown in following Figure 5 and Figure 6.

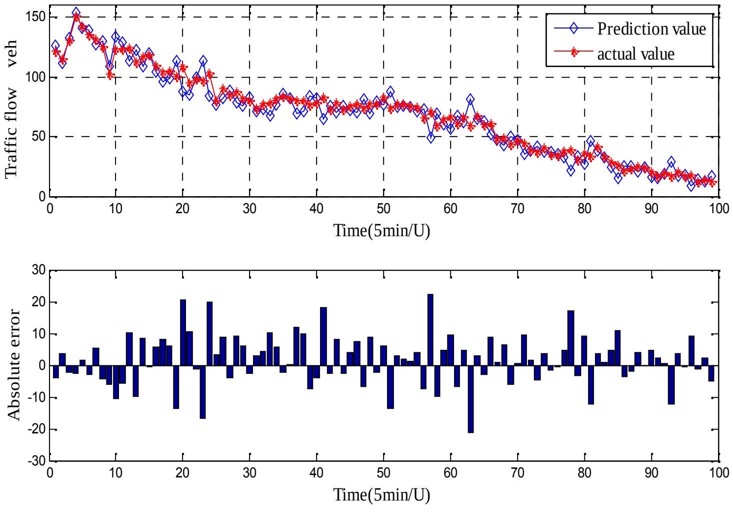

Figure 5. Comparison of Prediction Value and Detective Value- MPSR-LSSVM

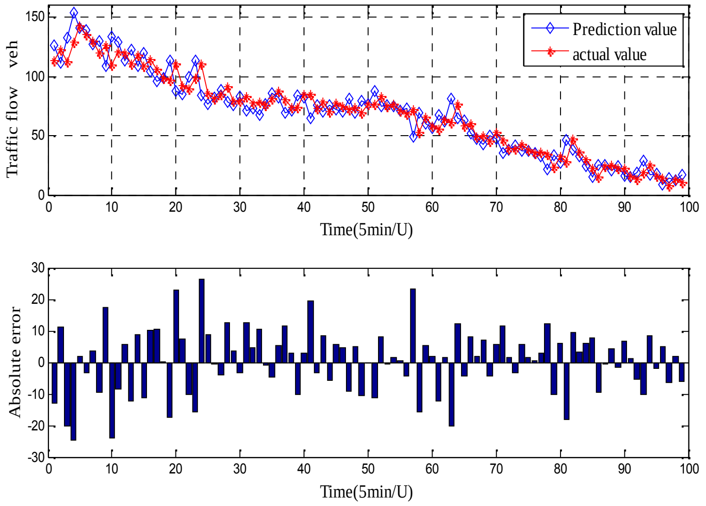

Figure 6. Comparison of Prediction Value and Detective Value- ASV-LSSVM

In order to compare the performance of the two models, this paper selects the following three indicators as a basis for evaluation. In these formulas below, f_i is the predicted value, y_i is the actual value, N is the number of predicted values. The comparisons of these two models are shown in Table 1 below.

1) Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{1}^{N}|f_i - y_i| \qquad (15)$$

2) Mean Relative Error

$$MRE = \frac{1}{N}\sum_{1}^{N}\left|\frac{f_i - y_i}{y_i}\right| \qquad (16)$$

3) Mean Square Error

$$MSE = \frac{1}{N}\sum_{1}^{N}(f_i - y_i)^2 \qquad (17)$$

Table 2. Evaluation Value of the Models

models	evaluation indexes		
	MAE	MRE	MSE
MPSR-LSSVM	5.3892	0.1136	50.5067
ASV-LSSVM	7.3348	0.1524	93.7449

From Table 2, it can be seen that the average absolute error (MAE), mean relative error (MRE) and mean square error (MSE) of MPSR-LSSVM model were less than ASV-LSSVM model's. Because correlation relationship exists in multivariate time series, and multivariate time series can eliminate noise of a single variable, therefore, it can reduce errors and improve the prediction accuracy.

5. Conclusion

Predicting traffic flow accurately is the basis of intelligent traffic control and guidance system, this paper presents a prediction model based on multivariate phase space reconstruction and LSSVM. Firstly, this model gets the embedding dimension and delay time of traffic flow, occupancy, and average speed time series, and then reconstructs multivariable phase space. Secondly, the phase points obtained after reconstruction are inputted, and the last traffic flow parameter came from the following phase point to be used as output. Finally, the SVM which be trained is adapted to realize short-term traffic flow prediction. This model differs from the previous models based on phase space reconstruction and support vector machine; it reconstructs traffic flows, occupancy, and average speed time series at the same time, while only traffic flow time series is restructured in earlier researches. Multivariate time series model may contain more dynamic information, and especially when there are some noises in the system, the model can filter out noise and improve the prediction accuracy. Experimental results show that the model can effectively predict short-term traffic flows. However, the model still has limitations, such as the dimension obtained after reconstructed is too large, this will be the direction of future research.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (No. 51208538)

References

[1] Hong WC, Dong YC. Forecasting Urban Traffic Flow by SVR with Continuous ACO. *Applied Mathematical Modelling*. 2011; 35: 1282~1291.

[2] Guo M, Sun ZQ. Research on Short Time Traffic Flow Forecasting Method. *Application Research of Computers.* 2008; 25(9): 2676~2678.

[3] Yang ZS*,* Wang Y. Short-term Traffic Flow Prediction Method Based on SVM. *Journal of Jilin University (Engineering and Technology Edition)*. 2006; 36(6): 881~884.

[4] Li MW, Hong WC. Urban Traffic Flow Forecasting Using Gauss-SVR with Cat Mapping, Cloud Model and PSO Hybrid Algorithm. *Neurocomputing*. 2013; 99: 230~240.

[5] Nair AS, Liu JC. *Non-Linear Analysis of Traffic Flow*. IEEE Proceedings of the 4th Intelligent Transportation Systems. Oakland. 2001: 681~685.

[6] Li Y, Liu B. The Identification of Chaos in Traffic Flow Using Surrogate-data Technique. *Systems Engineering.* 2000; 18(6): 54~58.

[7] Guo CL, Yu LJ. Research on Urban Traffic Flow Prediction Based on LS-WSVM. Logistics Engineering and Management. 2012; 34(9): 95~98.

[8] Yang W, Gong JL. Traffic Flow Prediction Based on Phase Space Reconstruction and Least Squares Support Vector Machines. *Transportation Science & Technology*. 2010; 242(5): 78~80.

[9] Takens F. *Determing Strang Attractors in Turbulence.* Lecture notes in Math. 1981; 898: 361~381.

[10] Kim HS, Eykholt R. Nonlinear Dynamics, Delay Times, and Embedding Windows. *Physica D*. 1999; 127: 48~60.

[11] Suykens JAK, Van Gestel T. Least Squares Support Vector Machine. Singapore: World Scientific Press. 2002.

[12] Rosenstein MT, Collins JJ. A Practical Method for Calculating Largest Lyapunov Exponents from Small Data Sets. Physica D. 1993; 65: 117-134.