3rd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2018)

Short-term traffic flow forecasting based on SVR

Li Yuanyuan[1, a], Xu Weixiang [2, b]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[2]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[a]15120763@bjtu.edu.cn, [b]wxxu317@126.com

Abstract: This paper constructs a short-term traffic flow forecasting model based on SVR. First, the modified KNN algorithm is applied to achieve phase space reconstruction and get the input data of SVR. Then, the short-term traffic flow forecasting model is established. Finally, this model was tested and evaluation indexes of traffic flow model were analyzed using the open microwave data provided by the OpenITS system. The results were compared with neural network and conventional SVR model and it shows that the model has better prediction performance.

Keywords: Short-term Traffic flow forecasting; Phase Space Reconstruction; KNN; SVR.

1. Introduction

In the intelligent traffic system, short-term traffic flow forecasting is mainly based on real-time dynamic traffic flow data. At present, short-term traffic flow forecasting research can be roughly divided into the following categories: The prediction method based on linear theory. This method mainly contains several methods, such as History Average Model, Autoregressive Moving Average Model, Time Serial Model, Kalman Filtering Model, Maxium Lidelihood Formulation Model and Markov prediction and so on[1]. The prediction methods based on nonlinear theory mainly include neural network prediction methods, methods based on support vector machines, chaos theory, wavelet analysis, etc[2-5]. In order to obtain more satisfactory results, two or more models are often combined to take the advantages of various prediction models, rather than relying solely on a prediction model or method.

In this paper, we construct a short-term traffic flow forecasting model based on SVR. First, the modified KNN algorithm is applied to achieve phase space reconstruction and get the input data of SVR. Then, the short-term traffic flow forecasting model is established. Finally, this model was tested and evaluation indexes of traffic flow model were analyzed using the open microwave data provided by the OpenITS system. The results were compared with neural network and conventional SVR model and it shows that the model has better prediction performance.

2. The short-term traffic flow forecasting model

2.1 Phase Space Reconstruction Based on KNN

The traffic flow system can be seen as a huge complex network. There is a significant non-linear relationship between the nodes, so the traffic flow time series can be viewed as a chaotic process and there is an interaction of nodes between the cohesive forces outside the attractors and the repulsive forces within the attractors. Xu Yongjun[6] calculated the Lyapunov exponent to prove that the short-term traffic flow is a chaotic time series. In the short-term traffic flow prediction, the forecast itself depends on the traffic flow data of the forecast point and its neighboring points or follow-up points. Therefore, the traffic data of the predicted points can be estimated by establishing a model with the above related points.

Packard[7] proposed reconstructing the phase space using the method of coordinate delay, that is, treating the points in a certain delay time as one new dimension point, thereby embedding a phase space that is “equivalent” to the original system. Taken's theorem proves that a suitable embedding dimension exists when one-dimensional time series is infinitely long without noise. The reasonable selection of appropriate delay time t and embedding dimension m can recover chaotic attractors in this embedding dimension space.

Assuming $\{x(t), t = 0, 1, 2, \dots, n)\}$ is a time series, the phase point of the m-dimensional phase space is:

$$\begin{aligned} X_1 &= \left(x(1), x(1+t), \dots, x(1+(m-1)t)\right) \\ X_2 &= \left(x(2), x(2+t), \dots, x(2+(m-1)t)\right) \\ &\dots\dots \\ X_N &= (x(N), x(N+t), \dots, x(N+(m-1)t)) \end{aligned} \tag{1}$$

In equation (1), m is the embedding dimension, t is the delay time, and N is the total number of phase points, and $N = n - (m-1)t$.

The KNN algorithm belongs to the supervised learning algorithm. There is a well-defined sample data, which have labels. In this paper, we will modify the criteria of the classifier. After given delay time t and embedded dimension m, we find the Euclidean distance between the samples in the high-dimensional space and select the k nearest distances, which can be called k nearest neighbors. In order to improve the accuracy, the corresponding method to reduce the amount of calculation is to set a k value with the lowest error rate: $k = \sqrt{N}$[8]. Due to uncertainty of the delay time and the embedded dimension, the value of k is not the same with different delay time and embedding dimension. Finally, we can choose the the most interdependent delay time t and embedding dimension m in the k nearest neighbors. The specific process how to obtain the delay time t and embedding dimension m by KNN algorithm, is shown in Fig. 1.

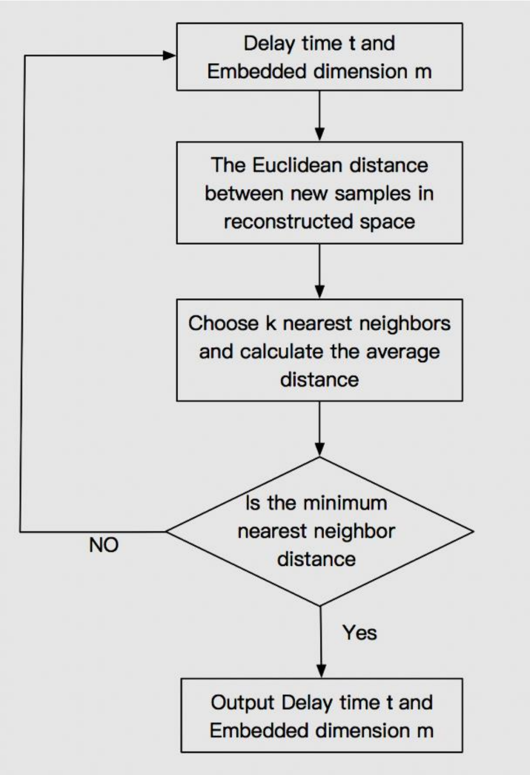
Delay time t and Embedded dimension m
The Euclidean distance between new samples in reconstructed space
Choose k nearest neighbors and calculate the average distance
Is the minimum nearest neighbor distance
NO
Yes
Output Delay time t and Embedded dimension m

Fig. 1 the flowchart for delay time and embedding dimension by KNN algorithm

The display of different distances with different delay time and embedded dimension is shown in Fig. 2. With the increasing t_m, which is the combination of t and m, the distance between the sample data gradually becomes longer, and the fluctuation occurs in the middle due to the difference t_m. By comparison, a reasonable delay time t and embedding dimension m are finally obtained by $t = 2min$ and $m = 15$.

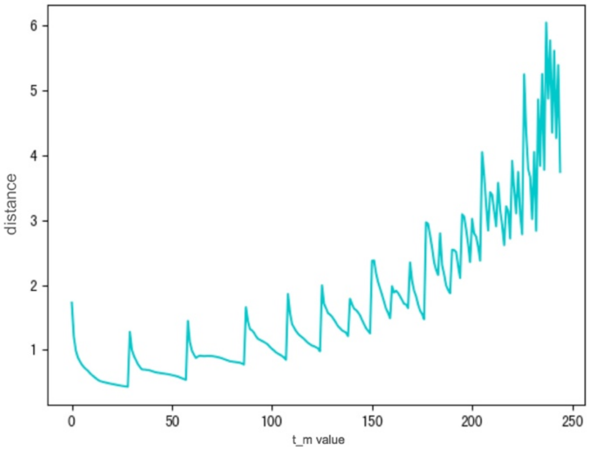

Fig. 2 the display of different distances with different delay time and embedded dimension

2.2 Support Vector Machine Regression

Support vector regression(SVR) is an improvement and application of SVM in regression. It can be understood as follows: When solving the linear problem, keep the point farthest from the regression line far from the regression line, so as to achieve the purpose that the sample data can be gathered in the same regression line. When solving a nonlinear problem, applying a kernel function to raise the dimension, and constructing a linear decision function in a high-dimensional space to achieve linear regression.

Assume that the sample set is $T = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n,\ y_i = \{-1, +1\},\ \mathrm{i} = 1, 2, \dots \mathrm{n}\}$. The support vector is the point who is far from the regression line, and the margin is distance between the support vector and the regression line. The selection criterion of the optimal regression line is the shortest margin. In the paper, Gaussian kernel function is used in SVR model and the regression line can be expressed as:

$$f(x) = \omega^{\mathrm{T}} \emptyset(x) + b \tag{2}$$

And the objective function is:

$$\begin{aligned} \min \quad & \|\omega\|^2/2 + C * \textstyle\sum_{i=1}^{l} (\xi_i + \xi_i^*) \\ s.t. \quad & y_i - (\omega^{\mathrm{T}} \emptyset(x_i) + b) < \epsilon + \xi_i \\ & (\omega^{\mathrm{T}} \emptyset(x_i) + b) - y_i < \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n \end{aligned} \tag{3}$$

In formula (3), C represents the penalty parameter of the error, ξ_i, ξ_i^* represent relaxation factors and ϵ represents the acceptable error. To solve this, we need to introduce Lagrange functions and factors α_i and α_i^*. Then, the optimal regression line can be determined as follows：

$$f(x) = \textstyle\sum_{i=1}^{n} (-\alpha_i + \alpha_i^*) K(x_i, x) + \mathrm{b} \tag{4}$$

In formula (4), $\mathrm{K}_{ij} = \emptyset(x_i)^{\mathrm{T}} \emptyset(x_j)$. Besides, the margin positive should equal to the margin negative so that b can be expressed as:

$$b = -(max_{i:y=-(\epsilon+\xi_i)} + min_{i:y=\epsilon+\xi_i})/2 \tag{5}$$

2.3 Model Construction

Assume that the time series is $\{x(t), t = 0, 1, 2, \dots, n)\}$, X_i is the i phase point of the phase space, t is the delay time, and the state transfer equation is:

$$X_i = \left(x_i, \dots, x_{i+(m-1)t}\right) \tag{6}$$

In this paper, we assume that the impact points include two parts, which are located in the upstream of the prediction point. The input and output of the SVR model based on phase space reconstruction can be expressed as:

$$input = \begin{cases} X_1 = \left(x_1, \dots, x_{1+(m-1)t}, x'_1, \dots, x'_{1+(m-1)t}\right)^{\mathrm{T}} \\ X_2 = \left(x_2, \dots, x_{2+(m-1)t}, x'_2, \dots, x'_{2+(m-1)t}\right)^{\mathrm{T}} \\ \dots \\ X_n = \left(x_n, \dots, x_{n+(m-1)t}, x'_n, \dots, x'_{n+(m-1)t}\right)^{\mathrm{T}} \end{cases} \tag{7}$$

$$output = \begin{cases} \mathrm{Y}_1 = x''_{1+(m-1)t} \\ Y_2 = x''_{2+(m-1)t} \\ \dots \\ Y_n = x''_{n+(m-1)t} \end{cases} \tag{8}$$

In formula (7), x and x' represent the upstream point of the prediction point, and x'' represent the prediction point. The flowchart of the model is shown in Fig. 3.

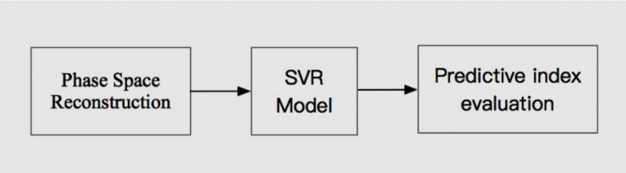

Fig. 3 the flowchart of the SVR model based on Phase Space Reconstruction

3. Experiment and comparative analysis

3.1 Experiment data

The openITS system is a public service platform based on the Internet that can realize access and sharing of relevant open datasets in the field of intelligent transportation [9]. In this paper, the open microwave data of the Huangke intersection in the demonstration area of Hefei City is applied, and the data collection time is from 6:30 to 9:30 and from October 11, 2016 to October 15, 2016. We take one lane as the prediction point and its two upstream lanes as the input.

3.2 Result and comparative analysis

In the SVR model, the more important parameters are penalty factor C, kernel function parameter γ, and error ϵ. In this paper, 10-fold cross validation is used. In the experiment, 3/4 data is used to train the model, then we can gain the result: $C = 5$, $\gamma = 0.0001$, and $\epsilon = 0.1$.

Then, we apply a three-layer BP neural network and the conventional SVR to compare analysis, and the trends of the predicted and measured values are shown in Fig. 4. It can be seen that all three models can fluctuate according to the actual data.

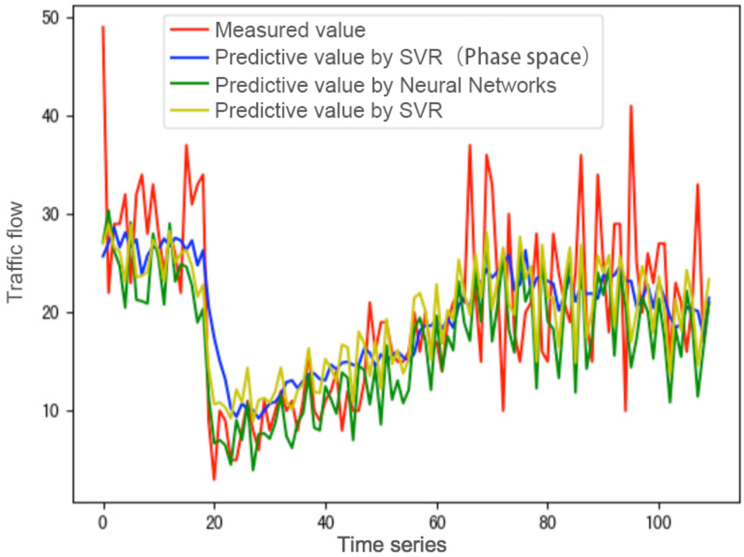

Figure 4 The curve between the measured value and the predicted value

In order to describe the forecasting abilities of the models, we introduce four evaluation indexes of traffic flow model: the mean square error(MSE), the mean absolute error(MAE), mean absolute percentage error(MAPE) and the equalization coefficient(EC). Then the calculated values of these indexes derived from above three models are shown in Table 1.

Table 1 The comparison between SVR and BP neural network model

Models	MSE	MAE	MAPE	EC
BP	59.15	5.66	0.29	0.81
SVR	48.62	5.17	0.30	0.83
SVR (phase space reconstruction)	38.89	4.64	0.29	0.85

As can be seen from Table 1, the SVR based on phase space reconstruction is superior to the other models in the three indicators of MSE, MAE and EC, and equal to the BP neural network in the MAPE index. Therefore, the SVR based on phase space reconstruction has good prediction performance.

4. Conclusions

This paper applies the open microwave data provided by the OpenITS system to predict the short-term traffic flow with the SVR model. When reconstructing the phase space, the algorithm based on KNN is used to obtain the embedding dimension and delay time. After that, the reconstruction result data is used as an input vector and input into the SVR model to predict the short-term traffic flow. The results show that the proposed model can effectively perform short-term traffic flow forecasting and is superior to the neural network and the conventional SVR model in most evaluation indexes of traffic flow model.

5. Acknowledgments

This work was financially supported by the National Natural Science Foundation of China General Projects Grant No. 61672002 and 61272029.

References

[1] Gao Hui, Zhao Jianyu, Jia Lei. Summay of Short Time Traffic Flow Forecasting Methods[J]. Journal of University Jinan(Sci&Tech), 2008, 22(1):88-94.
[2] Hasegawa M, Wu G, Mizuni M. Applications of nonlinear prediction methods to the Internet traffic[C]// IEEE International Symposium on Circuits and Systems. IEEE, 2001:169-172 vol. 2.
[3] Lin Jun, Ni Hong, Sun Peng. Adaptive Resource Allocation Based on Neural Network PID Control[J]. Journal of Xi'An Jiaotong University, 2013, 47(4):112-117.
[4] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.
[5] Xue J, Shi Z. Short-Time Traffic Flow Prediction Based on Chaos Time Series Theory[J]. Journal of Transportation Systems Engineering & Information Technology, 2008, 8(5):68-72.
[6] Xu Yongjun. Research for Short-term Traffic Flow Forecasting Method Based on Chaos and SVR[D]. Southwest Jiaotong University, 2011.
[7] Packard N H, Crutchfield J P, Farmer J D, et al. Geometry from a time series[J]. Physical review letters, 1980, 45(9): 712.
[8] Huang Juanjuan. Research and Improvement on Feature Selection and Classification Algorithms for Text Classification Based on KNN[D]. Xiamen University, 2014.
[9] http://www.openits.cn