A New Phase Space Reconstruction Method for Prediction of Public Transit Passenger Volume

XUE Xiang-hong[a] XUE Xiao-feng[b] XU Lei[c]

College of Computer Engineering, Jiangsu University of Technology, ChangZhou, China

[a]Xxh1226@jsut.edu.cn, [b]Xxf@jsut.edu.cn ,[c]Xl@jsut.edu.cn

Keywords: phase space reconstruction; support vector machine; passenger volume prediction; mutual information method

Abstract. This article proposes a new method for prediction of public transit passenger volume based on phase space reconstruction and support vector machine (SVM); applies mutual information method in calculating the optimal time delay of time series of public transit passenger volume; applies Cao's method in calculating the optimal embedding dimension; then calculates out the Lyapunov exponent, evidencing that passenger volume chaos phenomena exists. It establishes phase space reconstruction – support vector machine prediction model as well as determines the pairs of training sample for prediction of the real passenger volume. It is proved through actual examples that this method can predict passenger volume effectively.

Introduction

Accurate prediction of passenger volume of each stop in urban public transit is in favor of effective and timely public transit operation, for realizing both purposes of increasing economic benefits and meeting peoples' demands for taking public transportation means. As the key factor of public transit operation, the time series of passenger volume is featured with non-linearity and non-stability. Public transit passenger volume is related with many factors including climate, weather, holidays, major activity days, peak hours, station transfers, and ticket price etc, including both effects of certainty factors and random factors, having very complicated changes, for which traditional linear system has been unable to reveal its internal complicated features and laws. In non-linearity series studies, predication is often made with non-linearity model method such as artificial neural network etc[1]. Neural network can learn, drill and accumulate experience according to historical data, featured with adaptive ability; however, the results achieved are based on empiric risk minimization, for which sufficient and massive sample data are needed; there may be over-learning problems that lead to defect of poorer network promotion capability. This article proposes a new method for prediction of public transit passenger volume based on phase space reconstruction [2-3] and support vector machine (SVM) [4].

THEORY OF PHASE SPACE RECONSTRUCTION

Phase space reconstruction concept proposed by Packard and Takens etc introduces chaos theory into analysis of non-linearity time series. This theory believes that all the dynamical information needed for determining any system state is included in the time series of any variable for this system; the state trajectory achieved by embedding single-variable time series maintains the foremost characteristics of the original space state trajectory. Time series set with single variable is $\{ x (ti) , i = 1 , 2 \quad ,...,N\}$, in which N is series length, that means reconstructed phase space is:

$$Xi (t) = (x (ti) , x (ti +\tau) ,...,x (ti + (m - 1)\tau)) , i = 1 ,2 , ..., N -(m - 1)\tau \quad (1)$$

In which m is embedding dimension; τ is time delay; Xi is a point in phase space. According to Takens Theorem that if embedding dimension $m \geq 2d+1$, d is the dimension of dynamic system; then the reconstructed dynamic system and the original one is topologically equivalent. Therefore, the system state of the next moment can be acquired from the system's current state, thus acquiring the predication value for the next moment of the time series, which provides basis for predicting chaos time series [5].

In reconstructed phase space, selection of time delay τ and embedding dimension d is very important. It is reported from studies that if τ is too small, the system's dynamical characteristics will not be revealed; if it's too big, simple trajectory will be complicated and effective data points will be reduced as well. Samely, if d is too small, the embedding space will not contain dynamical system attractors, with which the system's dynamics characteristics cannot be revealed all around; contrarily if it's too big, it will not only reduce usable data length but also increase computation work volume, which may lead to increased prediction error.

A. Calculating Optimal Delay Time with Mutual Information Method

Methods of time delay selection mainly include auto correlation function and mutual information method. Since auto correlation function can just extract the linearity correlation of series space; while mutual information method is suitable for calculating the delay time of non-linearity time series. In this article, mutual information method is selected to determine reconstructed phase space delay time; delay time τ is the time corresponding to the min value point which is reached by mutual information function for the first time.

For time series {x(t1), x(t2) ,..., x(tn)}, delay time is set as τ, then time series is changed to be { xi+τ, i =1,2,...,n }， xk; the probability of occurrence in { xi,i=1,2,..., n } is P(xk)，xk+τ and is P(xk+τ) in { xi+τ, i=1,2,..., n }; joint probability of occurrence of xk and xk+τ in both series is P(xk, xk+τ), in which probability P(xk) and P(xk+τ) can be acquired through the occurrence frequency in the corresponding time series; joint probability P(xk,xk+τ) can be acquired through the value-corresponding grids on plane (xk, xk+τ), then mutual information function is:

$$I(\tau)=\sum_{k=1}^{N} P(x_k, x_{k+\tau})\ln\frac{p(x_{k,}x_{k+\tau})}{p(x_k)p(x_{k+\tau})} \qquad (2)$$

In which optimal delay time *τ* takes the first minimal value of mutual information function.

B. Calculating Embedding Dimension with Cao's Method

Methods for selecting embedding dimension generally include saturated embedding dimension and false nearest neighbors etc, in which false nearest neighbors is much widely used. This article adopts Cao's method[6] to determine embedding dimension; this method is an improved version of false nearest neighbors. For xi=(xi,xi+τ,…,xi+(m−1)τ)∈Rm,i=1,2,…,n

Definition:

$$a(i,m)=\frac{\left\|x^{(m+1)}(i)-x^{(m+1)}(j)\right\|}{\left\|x^{(m)}(i)-x^{(m)}(j)\right\|} \qquad (3)$$

In formula (3)：x(m +1)(i) is the ith phase point in the reconstructed phase space of dimension m +1； x(m +1)(j) is the nearest neighbor domain point of x(m +1)(i); ||·|| is Euclidean distance. For any phase point a(i, m), the average value is calculated with formula (4).

$$E(m)=\frac{n}{n-m\tau}\sum_{i=1}^{n-m\tau} a(i,m) \qquad (4)$$

E(m) just depends on embedding dimension m and delay time interval τ. In order to research the changes of phase space when embedding dimension is changed from m to m+1, it is defined:

$$F(m)=E(m+1)/E(m) \qquad (5)$$

In case F(m) gets saturated along increasing of m, now the value of m is the minimal embedding dimension of the reconstructed phase space.

PRINCIPLE FOR PREDICTION WITH SUPPORT VECTOR MACHINE

Support vector machine (SVM) is a new machine learning method proposed by Vapnik etc in 1995; it is based on statistical learning theory and minimal structure risk principle. Support vector machine is used to predict regression function; the basic principle is: mapping data input the space into another high-dimensional feature space in way of nonlinear mapping; then to carry out linear regression in this feature space[7].

PHASE SPACE RECONSTRUCTION─SVM PREDICTION MODEL

C. Determination of Series Chaos Characteristics

The key characteristic of chaos system is its sensitive dependency on initial conditions. Such dependency is measured with Lyapunov exponent, indicating the separation degree of two neighboring points in reconstructed phase space along with time evolution. A time series can be determined as chaos if the max Lyapunov exponent of it is over 0, meaning existence of attractors[8]. Lyapunov exponent can be calculated after delay time and embedding dimension are acquired. In this article, small data volume method is adopted to calculate the max Lyapunov exponent for the time series of public transit passenger volume to verify chaos characteristic existing in it.

D. SVM Prediction Model of Public Transit Passenger Volume

For given passenger volume time series $\{ x (t) , t = 1 , \dots , N\}$, phase space reconstruction method is adopted to convert it into a new data space with dimension m and time delay τ, which means: $Y (n) = [x (n - (m - 1)\tau) , \dots , x (n - \tau) , x (n)]$,In which $n \in [(m - 1)\tau, N]$, $Y (n)$ is the phase point after reconstruction; the value of m and τ can be acquired through above-mentioned methods. Prediction of passenger volume with after-reconstruction state vector to construct mapping (regression of prediction function) f : Rm→R, resulting in:

$$x (n + 1) = f (Y (n)) \qquad (6)$$

Suppose current moment is n, training data number is N, then training data can be expressed as: $(Y (n) , x (n + 1))$, $n = (m - 1)\tau, \dots , N – 1$. Determine training data with known sample series; apply support vector machine regression to train to acquire optimal model f. For prediction value of future moment $\hat{x}(t+1)$, take variable (m-1)τ in its reconstructed phase space as the input, predicting by applying support vector machine model acquired through training [9]。

CASE STUDY

In order to verify the prediction effect of the model generated by the algorithm, this article adopts passenger volume data collected from a bus route of a bus stand in a city. For influence of factors including peak hours, holidays etc and in order to measure data of pervasiveness, this article selects 4 days including Friday, Saturday, Sunday, Monday as data-collecting day; 2 data-collecting periods each day, one peak-hour period and one off-peak-hour period; totally 8 time periods. Each data collecting duration is 5 minutes; totally acquired 192 data of passenger volume time series.

First, adopt mutual information method to acquire the passenger volume delay time value 2; acquire embedding dimension 5 with Cao’s method, with these two values to carry out phase space reconstruction. Now use small-data method to calculate out the series’ max Lyapunov exponent λ=0.0497>0, indicating that the passenger volume time series is of chaos characteristics. 177 samples with dimension 5 are acquired after phase space reconstruction. The first 164 samples after reconstruction are taken as training samples; there are 13 test samples. Train these training samples with SVM regression. This article adopts Libsvm tool kit to realize regression prediction model. When using SVM, the most important thing is function selection; current common-use kernel functions include linear kernel function, multinomial kernel function, radial basis kernel function, and sigmoid kernel function. It is proved with experiment that compared with other kernel function, radial basis kernel function has less parameters as well as better performance[10]; therefore, radial basis kernel function is selected for this article. Selection of SVM model parameters is essential to the prediction result[11], including penalty factor c, insensitive factor ε and nuclear parameter γ. After determination of training sample set, finally select parameters C=58, γ=0.45, ε=0.00001 through model parameter tentative calculation. Finally, predict the tested data with the trained model, the prediction curve is as shown in Fig. 1.

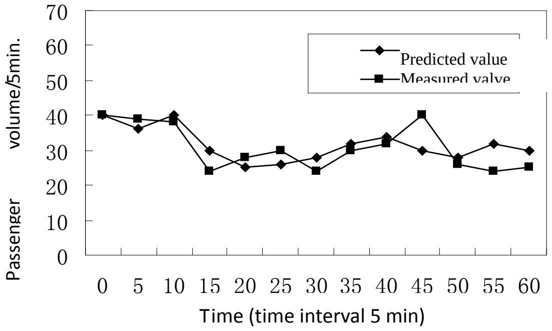

Fig. 1. Predicted Value of Passenger Volume Time Series

To verify the effectiveness of the method stated in this article, BP neural network model and single support vector machine model is established for prediction test. The predicted value of BP neural network & SVM regression model and the model of this article are taken as evaluation samples for comparison. In this study, model prediction performance will be evaluated with Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE); the smaller the index value, the more reasonable the model structure will be.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i-y_i)^2 \tag{7}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|\bar{y}_i-y_i\right| \tag{8}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\bar{y}_i-y_i}{y_i}\right| \tag{9}$$

In which $\bar{y}_i$ stands for model predicted value; yi is the value observed actually; n is the total quantity of prediction samples.

This article adopts three-layer-structured BP neural network prediction model, with S-shape logarithmic function Logsig() as transfer function of the output-layer neuron and S-shape tangent function Tansig() as transfer function for the hidden layer. Establish BP neural network prediction model with MATLAB neural network tool kit, with the first 170 data in above public transit passenger volume time series as drilling samples and the last 22 data as prediction samples. Single SVM regression prediction model is similar with the prediction model of this article; the difference between them is that no phase space reconstruction is carried out regarding passenger volume time series. The first 172 data of public transit passenger volume time series are taken as training samples and the last 20 data as prediction samples. The index results of prediction with the three models are as shown in the following table.

It is concluded from the comparison analysis of Table I, prediction mode (the model generated in this article) utilizing phase space reconstruction-support vector machine regression is of less errors, compared with that utilizing BP neural network and support vector machine regression, it has better prediction effect; therefore, it is more suitable for prediction of public transit passenger volume.

Conclusion and discussion

On the basis of studying that the time series of public transit passenger volume have chaos characteristics, this article establishes phase space reconstruction-SVM prediction mode by combining phase space reconstruction theory with support vector machine regression, so as to predict public transit passenger volume. It is proved through experiment comparison that the prediction model of this article is more precise, more suitable for passenger volume prediction. However, this article is of limitations, it is just designed for one-dimensional factor (passenger volume). In next studies, multi-dimensional reconstruction theory will be utilized to handle Multi-dimensional factors, so as to achieve more precision with this prediction model.

TABLE I COMPARISON OF FORECASTING PRECISION EVALUATION INDEXES OF THREE MODELS

Model type	MSE	MAE	MAPE/%
BP neural network model	78．563 93	6．547 73	6．312 5
SVM regression model	47．632 71	5．346 54	5．988 2
Regression model based on phase space reconstruction-SVM	22．374 75	3．132 56	2．734 3

References

[1] JIANG Ping, SHI Qin, CHEN Wu-wei. Prediction of passenger volume based on Elman type recurrent neural network [J]. Journal of Hefei University of Technology(Natural Science),2008,31(3)：340-342.

[2] Packard N H,Crutchfield J P,Farmer J D, et al. Geometry From a Time Series[J]. Physical Review Letters(S0031-9007),1980,45(9)：712-716.

[3] Takens F. Detecting Strange Attractors in Turbulence[J].Lecture Notes in Mathematics. Berlin：Springer-Verlag,1981 (898)：366-381.

[4] YU Jun-mei, YANG Yi. Real-time Passenger Volume Forecasting in Bus Stops Based on SVR [J]. Journal of Yantai College of Education,2008,14(1)：79-83.

[5] Ma Hongguang, Li Xihai, Wang Guohua. Selection of Embedding Dimension and Delay Time in Phase Space Reconstruction [J]. Journal of Xi＇an Jiaotong University, 2004, 38(4): 335-338.

[6] Cao Lianyue. Practical method for determining the minimum embedding dimension of a scalar time series [J].Physica D ,1997 ,110(5):43-50

[7] ZHANG Xuegong, introduction to statistical learning theory and support vector machines [J]. Acta Automatica Sinica,2001,01:36-46

[8] FAN Qian, HUA Xianghong. A Novel Method for Forecasting Landslide Displacement Based on Phase Space Reconstruction and Support Vector Machine [J]. Geomatics and Information Science of Wuhan University, 2009 ,34 (2) :248-251

[9] LUO Yun-qian, XIA Jing-bo, WANG Huan-bin. Application of Chaos-support Vector Machine Regression in Traffic Prediction [J]. Computer Science, 2009 ,36 (7) :244-246

[10] Hsu CHih-Wei,Chang Chih-Chung ,Lin Chih-Jen. A practicalguide to SVM classification [EB/OL].[2008-07-03.http: // www. csie. nt u. edu. tw/ ～cjlin/ papers/ guide/ guide. Pdf

[11] LIU Jing-xu, CAI Huai-ping, TAN Yue-jin. Heuristic Algorithm for Tuning Hyperparameters in Support Vector Regression [J]. Acta Simulata Systematica Sinica,2007,19(7):1540-1543