Recognizing Discrepant Traffic Data Based on Least Square Support Vector Machine

Chen Yao, Jiuchun Gu
College of Traffic and Transportation
Ludong University Yantai, Shandong
264000, China
yc@ibhsedu.com

Journal of Digital Information Management

ABSTRACT: *The real traffic databases are highly susceptible to noisy, missing, and inconsistent data. As reason of processing the discrepant data, a model is built based on square support vector machine which is good at dissolving the problems such as small samples, nonlinear and pattern recognition. Utilizing of the SVM, the inaccurate data can be detected by calculating the difference value between the real and prediction data. Comparing with the method of threshold theory, the model is proved to be better for the accurate data detection online and database cleaning.*

Subject Categories and Descriptors
I.1.2 [Algorithms]: Least square method; **H.2.7 [Database Administration**]: Data detection

General Terms: Prediction Model, Least Square Method

Keywords: Traffic Flow, Discrepant Data, Data Quality Control, Support Vector Machine, Least Square

Received: 1 June 2015, Revised 3 July 2015, Accepted 10 July 2015

1. Introduction

Traffic flow data is the basis for carrying out all transport-related research. The quality of traffic flow data makes great impact on the effect of the implication of traffic model [1]. In reality, due to the detection of equipment failure, communication interruption and adverse environmental factors, errors|, missing and lots of noise often appear in the traffic flow data acquisition system during the data collecting work. Therefore, quality control must be made on the traffic flow raw data, to detect and repair the discrepant data in order to make sure the reliability and accuracy of the application of traffic model.

Currently, many experts and scholars did lots of research on the detection and repairment of traffic discrepant data. Pei Yulong and Ma Ji made some comparison between two traffic discrepant data detection methods which were based on variable threshold and traffic flow theory, focusing on introducing the steps of threshold limits selecting the real-time traffic data [2]. Based on the cause of discrepant data, Jiang Guiyan and Jiang Longhui, etc. designed a set of methods which could identify the dynamic traffic data, but this method was still based on threshold limits [3]. Geng Yanbin defined the ITS data quality control and established a data control algorithm [4]. Qin Ling, Guo Yanmei, etc put forward traffic data testing and pre-testing methods including data cleaning, data repair and data smoothing, aiming at dealing with problems existing in the cross-section traffic testing data. [5] Although the above studies focus on different points, but they use the same basic idea, which is the threshold theory, traffic flow theory or a combination of both. In practice, such an approach is simple and easy to implement, but the choice of the threshold model needs to be determined by experience, which greatly affect the range and capabilities of the model on testing the discrepant data, particularly in dealing with massive traffic data, this defect is more prominent.

SVM (support vector machine, SVM) was first proposed by Vapnik. It's a machine learning method, which has the characteristics of fast learning, global optimization and generalization. It has made advantage in solving the small sample, nonlinear and high dimensional pattern recognition problems [6]. Here, I will use this method in testing the discrepant data in traffic timing order, to construct time order of nonlinear regression prediction model. Identify the outliners in data testing by comparing the residuals between the estimated and actual values.

2. Support Vector Machines

Support vector machines use least-squares linear system as a loss function [7]. Change the quadratic programming problem into linear programming in the standard SVM algorithm. Therefore, when dealing with massive data, we need least computational resources and get fast speed. Suppose a given sample $S = \{(x_1, y_1),...,(x_l, y_l)\} \in (X \times Y)^l$ $X = R^n, Y = R$, X、Y represent input data sets and output data sets respectively. In order to realize the regression estimation under the initial right space, we construct an optimized question as follow:

$$\min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{l=1}^{n} e_k^2 \qquad (1)$$

$$s.t.\ y_l = w^T \varphi(x_l) + b + e_l,\ l = 1,...,n$$

in this formula, J represents lost function, w represents weight vector, e_l represents error variable, b is partial term, γ is adjustable parameters which is for controlling the level of punishment that beyond the error of the sample, $\varphi(\ )$ is the nuclear space mapping function, which maps a sample from the original space R^n to R^{nk}.

In order to change the constrained optimization problem into the unconstrained optimization problem, we introduce the Lagrangian function:

$$L(W,b,e,a) = J(w,e) - \sum_{l=1}^{n} a_l \{w^T \varphi(x_l) + b + e_l - y_l\} \quad (2)$$

in this formula, a_l is the Lagrange operator. According to the nonlinear optimal planning KKT conditions, we try to get w、b、e_l、y_l partial derivatives in the formula (2) respectively and set the value as zero, the matrix equation is as followed

$$\begin{vmatrix} 0 & 1 & \cdots & 1 \\ 1 & \varphi(x_1)^T \varphi(x_1) + 1/\gamma & \cdots & \varphi(x_1)^T \varphi(x_n) + 1/\gamma \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \varphi(x_n)^T \varphi(x_1) + 1/\gamma & \cdots & \varphi(x_n)^T \varphi(x_n) + 1/\gamma \end{vmatrix} \bullet \begin{vmatrix} b \\ a_1 \\ \vdots \\ a_n \end{vmatrix} = \begin{vmatrix} 0 \\ y_1 \\ \vdots \\ y_n \end{vmatrix} \qquad (3)$$

through the Mercer theory, the nuclear functions $K(\ ,\ )$ make the $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ work, so it can be used to construct least support vector machines' decision function which is for regression estimation.

$$y(x) = \sum_{l=1}^{n} a_l K(x, x_l) + b \qquad (4)$$

Among them, the formula (3) solves the parameters a、b. As for kernel function, you can choose any symmetric function which meet the Mercer theory, commonly use sigmoid function, RBF function, and polynomial kernel function and so on.

3. Detection of Abnormal Data Which is Based on Least Square Support Vector Machine

3.1 Phase Space Reconstruction

SVM can only act on the vector set, so when we use it for time series abnormal data detection, time series data must first be converted into vector. According to phase space reconstruction theory, for the time series $x(t), t = 1, ..., n$, select the appropriate time lag τ and embedding dimension m, which could build a multi-dimensional vector space:

$$X = \{X(t) \mid X(t) = [x(t), x(t+\tau),..., x(t+(m-1)\tau)]\}$$

$t = 1, ..., n - (m-1)\tau$, making the reconstruction phase space in the embedded space "*trajectory*" and the original system dynamics equivalent under the meaning of diffeomorphism [8].

3.2 Principle and Steps of Abnormal Data Detection

Using support vector machine regression algorithm to detect time series abnormal data based on two considerations: First, support vector regression algorithm has good smoothness in the structural risk function. It does not tend to eliminate the individual regression error but to consider the smoothness of the return curve overall, so it's easy to observe because there is a significant distance between the outliers and the predictive value of regression function [8]. Second, the abnormal data and other measurements are not from a unified model, it is most likely to fall outside the loss function $[-\varepsilon, \varepsilon]$ (ε is non-sensitive areas). According to the KKT conditions, sample points' parameters value which are not in the non-sensitive areas must satisfy the. $a_l = 1/\gamma$ Based on support vector machine regression estimation theory, the basic detection steps and flow chart which are for summarizing the abnormal data are as follows:

(1)Extract the traffic data, using phase space reconstruction technique; change the time series data into vector data.

(2)Using the training data to set up the regression estimated model which is based on the least squares support vector machine.

(3)Select all the sample points which satisfy the conditions $a_l = 1/\gamma$ and calculate the residuals $E_i = |y_i - \hat{y}_i|$, y_i is the predicted value and is the measured value.

(4)Based on the accuracy requirements of traffic data collection, we introduce a constant $\lambda > 0$ as a criterion. If $E_i \geq \lambda$, the sample *i* is considered as abnormal values.

4. Case Study

Using video to capture the traffic data of the north second cross-section A, B, C in the First Ring Road in Chengdu. Traffic flow direction from west to east, study section shown in Figure 2. Data acquisition time is from 7:30 am to 10:30 am, July 21, 2008. Cycle time is Δt =2min intervals, totally get 90 data sets.

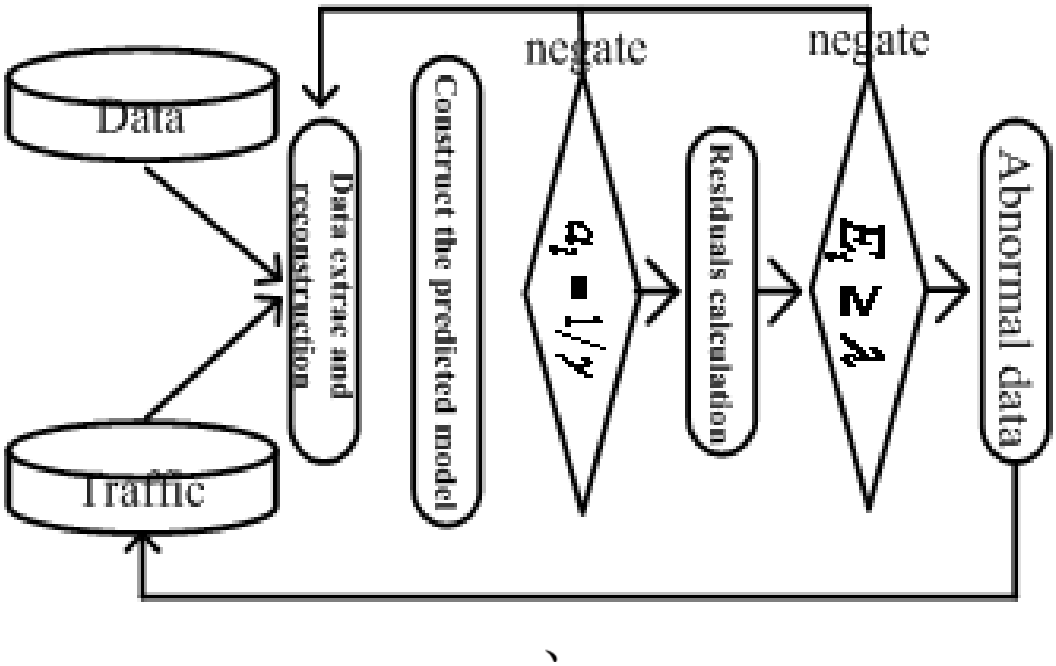

Figure 1.Traffic abnormal data detection flow chart based on least squares support vector

Define variables $Q_0(t)$ $Q_1(t)$、$Q_{-1}(t)$ are the total traffic flow which cross the cross-section A, B, C in the time period $t-2$ to t. T_0、T_1、T_{-1} are the intersection average delay which are connected to the section A, B, C.

Select, $Q_0(t)=\{t|t-2\Delta t,t-\Delta t,t\}$, $Q_1(t)=\{t|t-2\Delta t,t-\Delta t,t\}$, $Q_{-1}(t)=\{t|t-2\Delta t,t-\Delta t,t\}$ and T_0、T_1、T_{-1} for the input data set and $Q_0(t+\Delta t)$ for the output data, to construct the regression estimation model based on least squares vector the model kernel function is the RBF radial basis machine,function.

To verify the validity of the mentioned model in abnormal data detection, we add abnormal data in the original sample manually to make $Q_{10i}' = Q_{10i} + \eta \quad i = (1,2,\cdots 9)$, set constant η = 15. The first 50 data sets are put into the model to start training, to determine the parameter values γ = 650,ε = 0.0015, λ=12, then use the remaining 40 sets for validation, which $E_{10} = 25$, $E_{20} = 23$, $E_{30} = 25$, $E_{40} = 25$ are beyond the residual criterion, so it is easy to determine the sample i =10、20、30、40 is the abnormal data points, detection results shown in Figure 3.

Using the abnormal data fault identification method described in reference [3] to process the same data,

Figure 2. Figure of study section

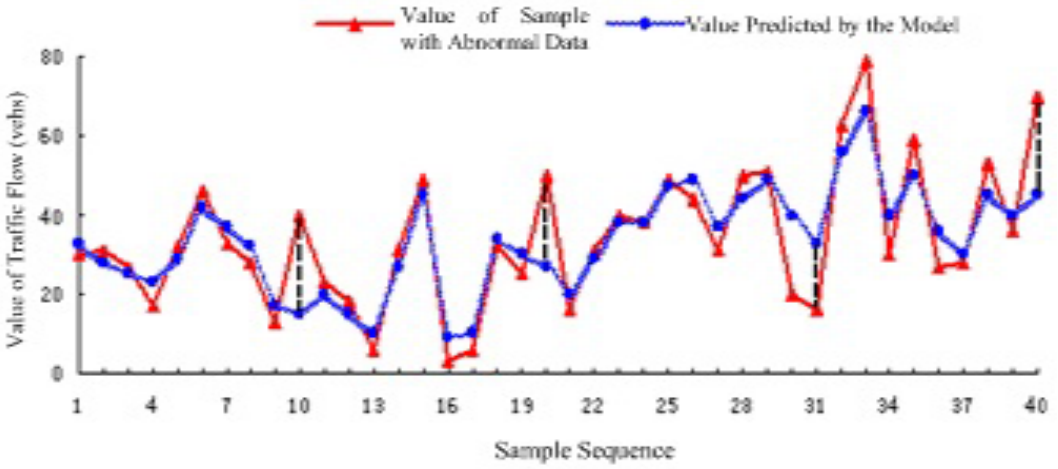

Figure.3 Model application renderings

determine the parameters'maximum flow value is *Qmax* = 82 *veh*/(2 *min*), we get the abnormal data point i =14020026033040. Through the comparison, we can see that this model is more efficient in practical applications for the detection of abnormal data traffic flow.

5. Conclusion

This article aims to construct least squares support vector machine regression model to test the abnormal traffic data, an example test and a comparison with the traditional method confirm the validity of the model. Taking into account the actual capacity of computing devices, when we use the model for multi-source, heterogeneous traffic data for online abnormal detection, how to exclude the interference between the data is for further research.

Support vector machine as a new data mining tools has been applied to many fields [9,10,11], which the core function is the key factor that constrain the development. How to select a kernel function reasonable and calibrate the model parameters are the research priorities in support vector machine.

References

[1] Jiang, Guiyan. (2004). Technologies and Applications of the Identification of Road Traffic Conditions. Beijing: China Communications Press.

[2] Yulong, Pei., Ji, Ma. (2003). Real-time Traffic Data Screening and Reconstruction. *China Civil Engineering Journal*, 36 (7) 78-83.

[3] Jiang, Guiyan., Gang, Longhui, Zhang., Xiaodong, et al. (2004). Malfunction Identifying and Modifying of Dynamic Traffic Data. *Journal of Traffic and Transportation Engineering*, 4 (1)121-125.

[4] Yanbin, Geng., Lei, Yu., Hui, Zhao. (2005). ITS Data Quality Control Techniques and Applications. *China Safety Science Journal*, 15 (1) 82-87.

[5] Ling, Qin., Yanmei, Guo., Peng, Wu , et al. (2006). Research on the Technology of Traffic Data Recognizing and Pre-processing. *Journal of Highway and Transportation Research and Development (Applied Technique)*, 11. 39-41.

[6] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Now York: *Springer*,138-145.

[7] Jiayuan, Zhu., Kaitao, Chen ., Hengxi, Zhang. (2003). Study of Least Squares Support Vector Machine. *Computer Science*, 30 (7) 157-159.

[8] Jinpei, Wu., Deshan, Sun. (2006) Modern Data Analysis. Beijing: China Machine Press, 272-279.

[9] Chapelle, V. N., Vapnik, O., Bousquet et al. (2002). Choosing multiple parameters for support vector machines. *Mach Learn*, 46 (1) 131–159.

[10] Carozza, M., Rampone, S. (2005). Towards an incremental SVM for regression, *In*: Proceeding international joint conference on neural networks, Italy, 405–410.

[11] Kowk, J.T. (2002). The evidence framework applied to support vector machine. *Trans Neural Network,* 11 (5) 1162–1173.