TELKOMNIKA, Vol.14, No.2A, June 2016, pp. 298~306
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v14i2A.4332

An Improved Double-layer K-nearest Neighbor Nonparametric Regression Method for Short-time Traffic Flow Prediction

Wang Cheng*[1], Pang Xiyu[2], Huang Guolin[3]
[1,2]School of Information Engineering, Shandong Jiaotong University, Shandong, Jinan, 250357, China
[3]Department of Intelligent Traffic Engineering, Yigou Software Technolog Co., Ltd. Shandong, Jinan, 25000, China
*Corresponding author, e-mail: wangcheng_1001@163.com[1], xiyupang@126.com[2], huanggl@163.com[3]

Abstract

In combination with the repeatability of the traffic flow state patterns, this article improved the k-nearest neighbor non-parametric regression method. To be specific, the neighbors were screened twice and the function based on state pattern recognition was introduced; moreover, the traffic flows in the past time and the traffic flows towards the related directions at both upstream and downstream crossroads were taken into account, so that the predictive ability of the proposed k-nearest neighbor non-parametric regression method can be improved. In addition, the final prediction results were output using the weighted average method of the reciprocal of the state pattern vector matching distance, so as to enhance the accuracy and real-time performance of the short-term traffic flow prediction.

Keywords*: Traffic flow; Double-layer; State pattern; Non-parametric*

1. Introduction

With the acceleration of urbanization and mechanization, the increasing car ownership, especially the rapid development of private cars, has led the ever-growing traffic flows and the worsening contradiction between urban traffic demand and supply. Therefore, the adoption of reasonable traffic demand control and road traffic control measures is the key to solving the urban transport problems, in which the accurate estimation of road traffic flows is the bottleneck in reasonable guidance, controlling and management the transportation. The prediction of traffic flow refers to the prediction of the traffic flow at the moment of $(t+\Delta t)$ based on the existing historical traffic flow data at the moment of t, with the prediction time generally smaller than 15 minutes. Only with the high-accuracy real-time traffic flow information, can the optimal driven routes be provided for the travelers with the use of modern information technology and therefore the goals of smooth network and high-efficient operation be achieved.

The domestic and foreign researchers have conducted many studies regarding the real-time traffic flow prediction. In terms of the parameters used in the prediction, the real-time traffic flow prediction methods can be classified into the parametric prediction methods and nonparametric prediction methods. With regard to the parametric prediction methods [1-4], some restrictive assumptions are made on the modeling data; moreover, these data can be fitted by the mathematical expressions with a finite number of parameters or the distribution of these data is known. Otherwise, the methods can be regarded as the nonparametric prediction methods. The parametric prediction methods mainly include moving averaging method, exponential smoothing method, time series method, Kalman filtering method and etc., while the nonparametric prediction methods include neural network, nonparametric regression method, the method based on wavelet decomposition and reconstruction and so on [5-7].

Using the parametric methods [8-12], some characteristics of the traffic system such as uncertainty, complexity and dynamic nature can be hard to be accurately simulated and represented. Nonparametric regression is a nonparametric prediction method applicable to the complex and dynamic systems with great certainties, and can be used for predicting the real-time traffic flow in short time. Currently, the single short-time traffic flow prediction methods all require the unique information characteristics and the specific application conditions, leading to poor prediction accuracy in dealing with the complex traffic flows. Moreover, in order to select

Received January 26, 2016; Revised April 28, 2016; Accepted May 17, 2016

the optimal method, a vast number of calculations should be performed before the prediction, which is unfavorable to short-term traffic flow prediction [13-16].

K-nearest neighbor nonparametric regression method now is a widely-applied nonparametric regression algorithm and exhibits a series of advantages such as parameter-free, small error ratios and favorable error distributions. More importantly, the improvements on search algorithm and the adjustment rules of parameters can make this method really satisfy the requirements in real-time traffic flow prediction.

In this article, the K-nearest neighbor nonparametric regression method was improved in the following three aspects. One is the adoption of two-layer screening of neighbors, i.e., the neighbors were only screened once using the traditional K-nearest neighbor nonparametric regression method while the neighbors were screened twice using the improved method, so that the prediction accuracy can be enhanced. Furthermore, the function based on state pattern recognition was introduced, i.e., the traffic flows in the past time and the traffic flows towards the related directions at both upstream and downstream crossroads were taken into account. Accordingly, the predictive ability of K-nearest neighbor nonparametric regression method was improved. Finally, the prediction results were output using the weighted average method of the reciprocal of the state pattern vector matching distance, so as to enhance the accuracy and real-time performance of the short-term traffic flow prediction.

2. Research Method

The flow chart of the proposed state-pattern-based short-time traffic flow prediction method is shown as figure 1, and the algorithm consists of the following steps:

(1) Establish the historical standard sample database;

(2) Construct the time-series-based traffic flow state vector;

(3) Calculate the difference between two adjacent elements in the state vector;

(4) Conduct the normalization processing on the differences and construct the traffic flow state pattern vector;

(5) Establish the state pattern matching distance for the evaluation of similarity;

(6) Evaluate the similarity between the current point and the point in historical database, conduct the preliminary screening of the points and select n neighbors;

(7) Construct the state vectors of the traffic flows at the current section and the traffic flows towards the related directions at upstream and downstream crossroads;

(8) Construct the weighted Euclidean matching distance;

(9) Select k neighbors;

(10) Output the prediction results using the neighbor non-parametric regression method based on the weighted average method of the reciprocal of the state pattern vector matching distance.

2.1. Establishment of Historical Standard Sample Database

Historical database is also referred to as the source case database. The quality of the prediction results mainly relies on the completeness of the historical database. As the traffic information acquisition technology advances, the information acquisition range expands and the information acquisition precision increases, which makes the acquisition of sufficiently high-quality historical data possible. The historical database should not only include the variation tendencies and typical rules of various traffic states, but also consider the fact that the redundant data can make the operation of algorithm too time-consuming. More complete data include more traffic flow states, so that the nearest neighbor can be found and more accurate prediction results can be obtained. However, too much data is also unfavorable for the searching of K-nearest neighbors and can cost too long time. As a consequence, the redundant data should be simplified.

Figure 2 displays the procedures of the establishment of historical standard sample database. Firstly, pre-processing was conducted on the historical data, the abnormal data were identified, the wrong data were rejected and the missing data were mended. Then, the redundant data were removed using the matching algorithm based on the Euclidean distance between the state vectors, and the historical standard sample database was established.

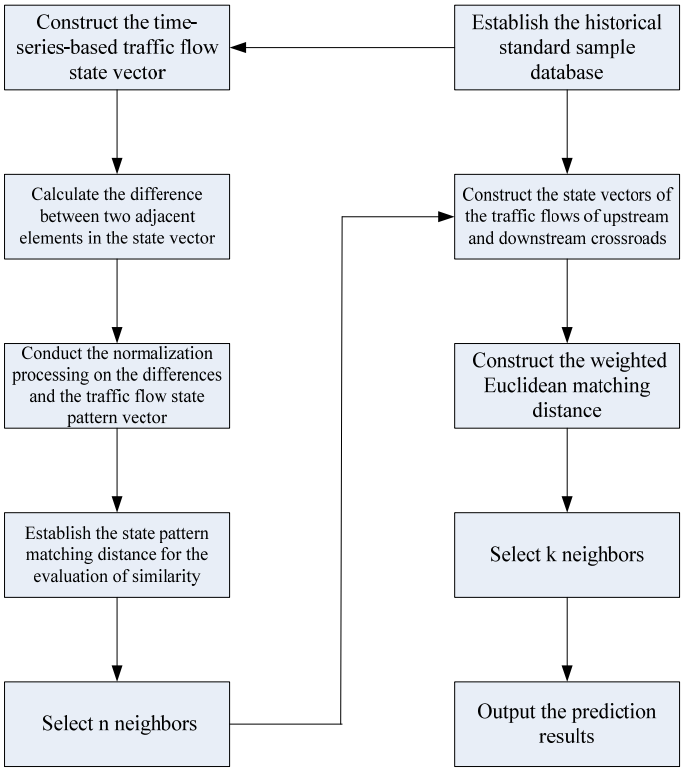

Figure 1. The flow chart of the improved traffic flow prediction method

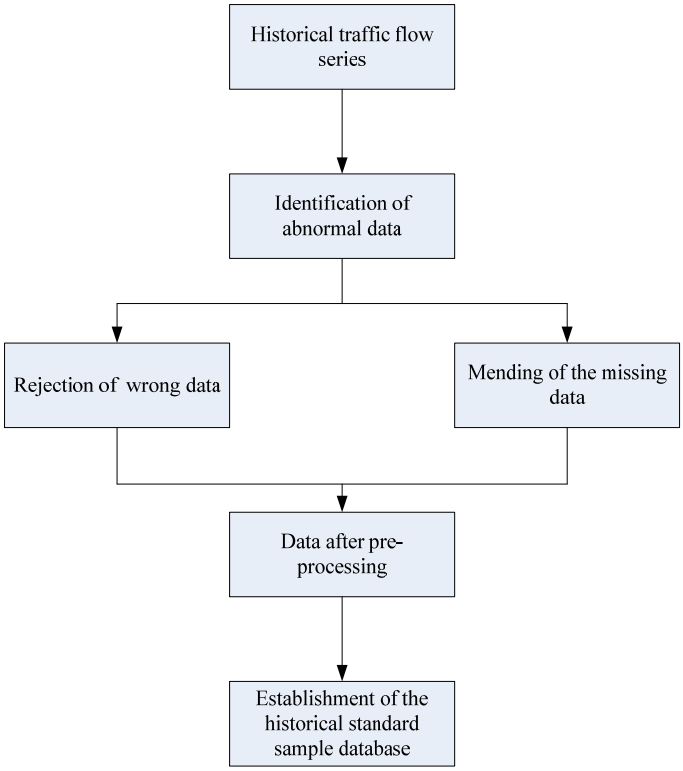

Figure 2. Flow chart of the establishment of historical standard sample database

2.2. Neighbor Nonparametric Regression Method Based on State Pattern Identification

(1) Preliminary matching and screening

In this article, the pattern matching and preliminary screening of the points in the historical database were conducted using the neighbor nonparametric regression method based on traffic flow state pattern vector, and the specific procedures of the algorithm are described below.

Firstly, the traffic flow time series was used as the traffic flow state vector M, and the formula can be written as:

$$M(t)=\left[q(t-l+1),q(t-l+2),\cdots,q(t)\right] \quad (1)$$

in which $M(t)$ denotes the traffic flow state vector at the moment of t at the current road; q(t-l+1), q(t-l+2) and q(t)denote the traffic flows at the moments of (t-l+1), (t-l+2) and t at the current road, respectively; and l denotes the dimensions of the state vector.

Then, the normalization was carried out on the difference between the traffic flows at the adjacent time series, which was adopted as the traffic flow state pattern vector M_d. The formula can be written as:

$$\begin{aligned} &M_d(t)=\left[r(t-l+1),r(t-l+2),\cdots,r(t-1)\right] \\ &r(i)=\frac{d(i)-\min(d(i))}{\max(d(i))-\min(d(i))} \\ &d(i)=q(i+1)-q(i) \end{aligned} \quad (2)$$

in which $M_d(t)$ denotes the traffic flow state pattern vector at the moment of t at the current road; $r(i)$ denotes the normalized difference between the traffic flows at the moments of (*i+1*) and *i*; $\max(d(i))$ and $\min(d(i))$ denote the maximum and minimum of the traffic flow differences at any two adjacent moments of (*t-l*+1), (*t-l*+2), ..., (*t-1*) and *t*, respectively; and $d(i)$ denotes the difference between the traffic flows at the moments of (*i+1*) and *i*, respectively ($t-l+1\le i\le t-1$).

Finally, the similarity between the state pattern at the current point and the state pattern at the historical point was calculated using Euclidean distance, and the points in the historical database were screened based on the similarity of state pattern. The set of the screened points was denoted as A. The specific formula can be written as:

$$d_{mh}=\sqrt{\left(r(t-l+1)-r_h(t-l+1)\right)^2+\left(r(t-l+2)-r_h(t-l+2)\right)^2+\cdots+\left(r(t-1)-r_h(t-1)\right)^2} \quad (3)$$

in which, d_{mh} denotes the matching distance between the state pattern at the current point and the state pattern at the point in historical database; r(*t-l*+1), r(*t-l*+2), ... and r(*t-1*) denote the normalized differences of the traffic flows at the adjacent moments of (t-l+3), (*t-l*+2), ..., (*t-1*) and t, respectively; and r*h*(*t-l*+1), r*h*(t-*l*+2), ... and r*h*(t-*1*) denote the normalized differences of the traffic flows at the corresponding moments in the historical database, respectively.

The matching distances of the state patterns between the current point and the point in historical database were sorted in an ascending order, and n nearest neighbors were picked out, with the set of neighbors denoted as $A=\{q(t_1),q(t_2),\cdots,q(t_n)\}$.

(2)Secondary matching and screening

The points in the historical database were matched and screened secondarily using the improved K-nearest neighbor nonparametric regression method.

Firstly, the traffic flows at the current section and the traffic flows towards the relation directions at the upstream and downstream crossroads were denoted as the state vector X, as shown in Figure 3. X can be calculated by:

$$X(t)=\left[v_1^u(t),v_2^u(t),\cdots,v_m^u(t),v(t),v_1^d(t),v_2^d(t),\cdots,v_j^d(t)\right] \quad (4)$$

in which, $X(t)$ denotes the traffic flow state vector at the moment of t at the current road; $v_1^u(t)$, $v_2^u(t)$, ..., $v_{\mathrm{m}}^u(t)$ denote the traffic flows at the upstream crossroad towards the related directions; m denotes the number of directions at the upstream crossroad; v(t) denotes the traffic flow at the moment of t at the current road; $v_1^d(t)$, $v_2^d(t)$, ..., $v_{\mathrm{j}}^d(t)$ denote the traffic flows at the downstream crossroads towards the related directions, and j denotes the number of directions at the downstream crossroad.

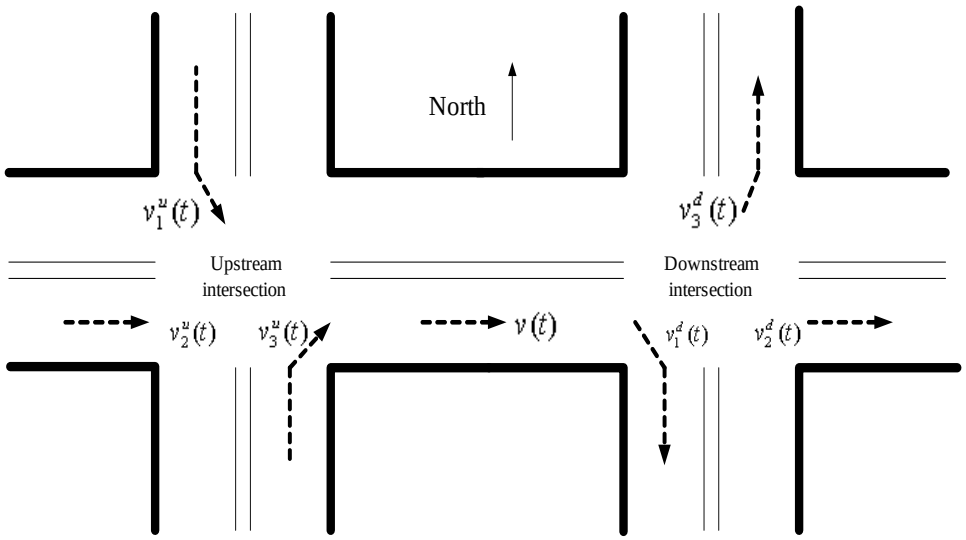

Figure 3. Illustration of a typical urban road network

Then, the weighted Euclidean distance method was used for evaluating the state similarity between the current point and the point in the set A, and the formula can be written as:

$$d_h=\sqrt{\begin{array}{l}a_1(v_1^u(t)-v_{h1}^u(t))^2+a_2(v_2^u(t)-v_{h2}^u(t))^2+\cdots+a_m(v_m^u(t)-v_{hm}^u(t))^2+b(v(t)-v_h(t))^2\\+c_1(v_1^d(t)-v_{h1}^d(t))^2+c_2(v_2^d(t)-v_{h2}^d(t))^2+\cdots+c_n(v_n^d(t)-v_{hj}^d(t))^2\end{array}} \quad (5)$$

in which , d_h denotes the matching distance between the current point and a point in A, $v_1^u(t)$; $v_2^u(t)$, ..., and $v_{\mathrm{m}}^u(t)$ denote the traffic flows at the moment of t at the upstream crossroad towards the related directions; $v_{h1}^u(t)$, $v_{h2}^u(t)$, ..., and $v_{\mathrm{hm}}^u(t)$ denotes the traffic flows at the point in A towards the related direction at the upstream crossroad; m denotes the number of directions at the upstream crossroad; v(t) denotes the traffic flow at the moment of at the current road; $v_1^d(t)$, $v_2^d(t)$, ..., $v_{\mathrm{j}}^d(t)$ denote the traffic flows towards the related directions at the downstream crossroad; $v_1^d(t)$, $v_2^d(t)$, ..., and $v_{\mathrm{m}}^d(t)$ denote the traffic flows at the moment of t towards the related directions at the downstream crossroad; j denotes the number of directions at the downstream crossroad, and $\{a_1,a_2,\cdots,a_m,b,c_1,c_2,\cdots,c_j\}$ denotes a set of weighted values ($a_1\in[0,1],a_2\in[0,1],\cdots,a_m\in[0,1],b\in[0,1]$, $a_1\in[0,1],a_2\in[0,1],\cdots,a_m\in[0,1],b\in[0,1]$, $a_1\in[0,1],a_2\in[0,1],\cdots,a_m\in[0,1],b\in[0,1]$ and $c_1\in[0,1],c_2\in[0,1],\cdots c_j\in[0,1]$).

Finally, the matching distances of the state patterns between the current point and the point in A were sorted in an ascending order, and n nearest neighbors were picked out, with the set of neighbors denoted as $B=\{q(t_1),q(t_2),\cdots,q(t_k)\}$.

2.3. Weighed Averaging on the Reciprocal of State-Pattern-Based Matching Distance

In this article, the prediction function was constructed using the weighted averaging on the reciprocal of the state pattern matching distance, and the traffic flow at the next moment was predicted by the most similar stated. In addition, the prediction results using the k-nearest neighbor nonparametric regression method. The specific formula can be written as:

$$\tilde{q}(t+1)=\sum_{i=1}^{k}\left(\frac{1}{d_{mh}(i)}\Big/d\right)q(t_i+1)$$

$$d=\sum_{i=1}^{k}\frac{1}{d_{mh}(i)} \quad (6)$$

in which , $\tilde{q}(t+1)$ denotes the predicted traffic flow at the next moment at the current crossroad, k denotes the number of the elements in the set B (i.e., the selected nearest neighbor points in the historical database), $d_{mh}(i)$ denotes the distance of the state pattern between the current point and the nearest neighbor points in the historical database, i.e., the points in the set B, and $q(t_i+1)$ denotes the traffic flow at the moment of (t_i+1) in the historical database.

3. Results and Analysis

The real-time monitoring data at Daminghu Road, Jinan, China, were used in the present experiments. In view of the traffic flow's periodicity, the monitoring time was set from 6th June, 2015, to 13th June, 2015, including 5 workdays and 3 holidays. Since there is generally less traffic at night, the nighttime traffic data have little practical value, and the monitoring time was set from 7:00 a.m. to 20:00 p.m. The sampling interval was set as 2 minutes, and totally 3120 original traffic flow data samples were collected. The data in the first seven days were used for the establishment of historical sample database while the data in the last day were used for verifying the constructed model.

The proposed traffic flow prediction method involves six parameters--*l, n, m, j, b and k*. Specifically, *l* denotes the dimension of state vector, which directly determines the prediction accuracy and the efficiency of the algorithm; n denotes the number of the points after the preliminary screening based on state pattern matching; and k denotes the number of the points after the secondary screening based on state pattern matching. The values of n and k can directly affect the prediction accuracy and efficiency of the algorithm, i.e., too large or too small values of n and k can both reduce the prediction accuracy.

The values of the parameters *l, n and k* were acquired based on the above-described experimental data. As shown in figure 4, the prediction error is significantly reduced as *l* increases from 2 to 6; however, as *l* continue to increase, the prediction error almost remains at a same level. On the other hand, the increase of *l* leads to an increasing calculation burden. The optimal matching number is at around 4, suggesting that the increase of matching number cannot improve the prediction results to a great extent and may lead to opposite effects. Therefore, *l* was set as 4 in practical applications.

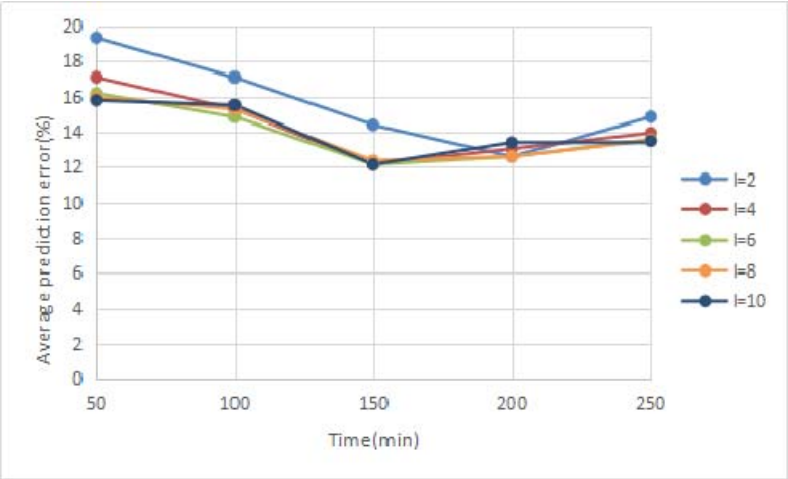

Figure 4. Effect of the value of l on the prediction accuracy

As shown in Figure 5, under the premise that the state vector and the prediction algorithm were finalized, as the value of *n* increases from 40 to 50, the prediction error decreases significantly; as the value of *n* further increases from 50 to 65, the prediction error gradually increases at a slow speed. The result suggests that the optimal matching number *n* should be set as 50. After the number of nearest neighbors was determined, the experiment was conducted on the secondary screening. *k* denotes the number of the points after the secondary screening based on state pattern matching. As shown in Figure 6, if the value of *k* is too large, the prediction function is over-smoothed, leading to the decline in prediction accuracy; however, if the value of *k* is too small, the effects induced by the accident factors were added, which can also affect the prediction accuracy. In this article, *k* was set as 9.

The numbers of the directions at the upstream and downstream crossroads were denoted as m and n. The upstream crossroad and the downstream crossroad of Daminghu Road both have three directions, i.e., $m = j = 3$. In practical applications, the traffic flow at the next moment at the current road is not only related to the traffic flow at this moment at the current road, but also connected with the traffic flows at this moment towards the related directions at the upstream and downstream crossroads. The effects of these factors have different weights. Based on the comprehensive analyses on the actual conditions of Daminghu Road, *b* was set as 0.5, and moreover, $a_1 + a_2 + \cdots + a_m = 0.2$, $c_1 + c_2 + \cdots + c_m = 0.3$.

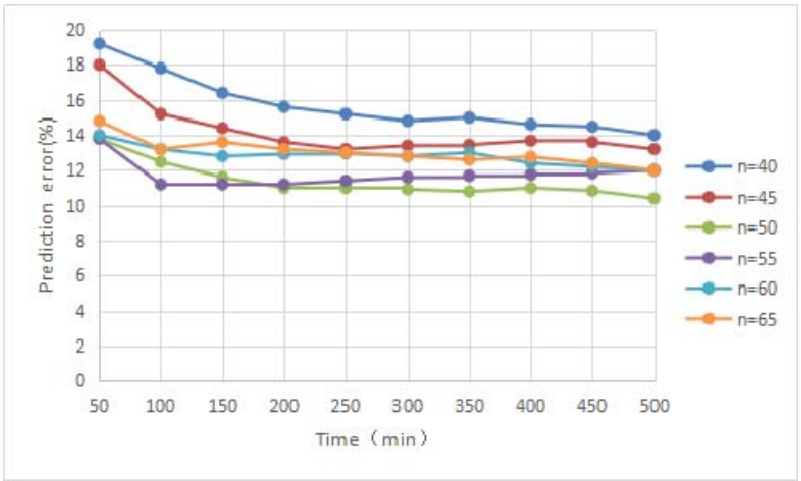

Figure 5. Effect of the value of n on the prediction accuracy

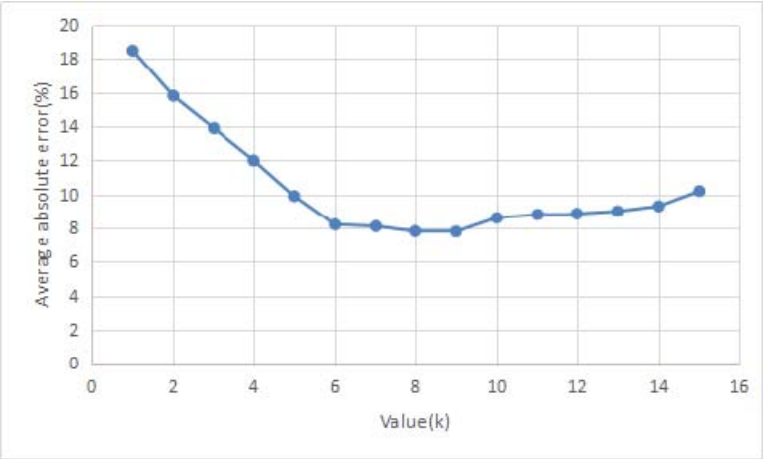

Figure 6. Effect of the value of k on the prediction accuracy

Finally, the results using the improved state-pattern-based K-nearest neighbor algorithm was compared with the results using the traditional K-nearest neighbor algorithm. The values of various parameters were substituted into the formulas for the prediction of the traffic flow at the next moment. The simulations were also conducted using Matlab. Figure 7 displays the simulation results, form which we can observe that the improved state-pattern-based K-nearest neighbor algorithm is superior to the traditional method in prediction accuracy. Conclusively, the proposed state-pattern-based K-nearest neighbor algorithm exhibits better prediction performances than the traditional method.

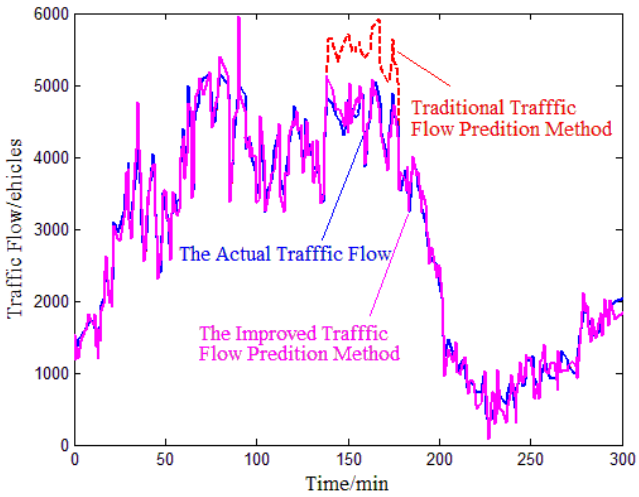

Figure 7. Comparison between the traditional method and the improved method

4. Conclusion

Short-time traffic flow prediction is an important part in intelligent traffic forecasting system. The traditional K-nearest neighbor nonparametric methods have not considered the mutual effects or the memorability of the actual traffic data. In this article, in view of the repeatability of the traffic flow state patterns, the traditional K-nearest neighbor nonparametric algorithm was improved. The double-layer neighbor nonparametric regression method was adopted, and the function of state-pattern-based recognition was introduced. Meanwhile, the traffic flows in the past time and the traffic flows towards the related directions at the upstream and downstream crossroads were taken into account, so that the predictive ability of the K-

nearest neighbor nonparametric regression method can be enhanced. Using the weighted averaging on the reciprocal of the state pattern matching distance, the final prediction results were calculated and output. Finally, according to the prediction results of the measured traffic flows, one can conclude that the improved two-layer K-nearest neighbor nonparametric regression method can enhance the accuracy and real-time performance in short-time traffic flow prediction, which is proved to be an effective short-time traffic flow prediction method. The prediction results can provide the evidence for the traffic management departments to conduct the related traffic guidance and control services, which is of great significance to traffic guidance and controlling.

Acknowledgements

The research work was supported by Shandong Provincial Natural Science Foundation, China Grant NO. ZR2014FL004 and A Project of Shandong Province Higher Educational Science and Technology Program Grant NO. J15LN12 and Shandong Province Independent Innovation and Transformation of Scientific Achievements Special Fund (New industries) Grant NO. 2015ZDXX0201A05 and Shandong Province Statistical Research Project Grant NO. KT15143 and Shandong Province National Economy and Social Informationization Development Soft Science Research Project Grant NO. 2015EI025 and Shandong Province Housing Urban and Rural Construction Science and Technology Project Plan --"Research on urban intelligent traffic flow forecasting system".

References

[1] Lima RM, Carvalho D, Vaccaro G, Scavarda LF. Industrial engineering and operations management–special issue. *International journal of industrial engineering and management (IJIEM)*, 2013, 4(3): 103-108.
[2] Ledoux C. An urban traffic flow model integrating neural networks. *Transportation Research Part C: Emerging Technologies*. 1997; 5(5): 287-300.
[3] Shi Yonghui, Bao Jun, Yan Zhongzhen. The Study of the hybrid intelligent algorithm in city road traffic flow forecasting. *Traffic Information And Security*. 2011; 4(4): 58- 61.
[4] Xu Yanyan, Zhai Xi, Kong Qingjie. Short-term prediction method of freeway traffic flow. *Journal of Traffic and Transportation Engineering*. 2013; 13(2): 114-119.
[5] Gao Wei, Lu Baichuan, Fu Tianli. Prediction of short-time traffic flow based on the temporal and spatial features and RBF neural network. *Traffic information and security*. 2011; 1(1): 16-19.
[6] Li Song, Liu LIjun, Xie Yongle. Chaotic prediction for short-term traffic flow of optimized BP neural network based on genetic algorithm. *Control and Decision*. 2011; 16(10): 1581-1585.
[7] Sheng JIN, Dian-hai WANG, Cheng XU, Dong-fang. Short-term traffic safety forecasting using Gaussian mixture model and Kalman filter. *Journal of Zhejiang University-Science A(Applied Physics & Engineering)*. 2013; 12(4): 51-53.
[8] Yu S, Wei YM, Wang KA. PSO-GA optimal model to estimate primary energy demand of China. *Energy Policy*. 2011; 42(12): 329-340.
[9] Liu Yan, Zhang Xun. The Application of the combination forecasting model in short-term traffic flow prediction. *Logistics Management*. 2010; 23(2): 15-19.
[10] Kamarianakis Y, Shen WL Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. *Applied Stochastic Models in Business and Industry*. 2012; 28(4): 297-315.
[11] M Abdar, SRN Kalhori, T Sutikno, IMI Subroto, G Arji. Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*. 2015; 5(6): 1569-1576.
[12] Hong, WC. Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm. *Neurocomputing*. 2011; 74(12): 2096-2107.
[13] Fan Na, Zhao Xiang-mo, Dai Ming. Short-term traffic flow prediction model. *Journal of Traffic and Transportation Engineering*. 2012; 12(4): 114-119.
[14] Qu Li, Lan Shiyong, Zhang Jianwei. Short-term traffic forecasting based on nonparametric regression and floating car data. *Computer Engineering and Design*. 2013; 34(9): 3298-3332.
[15] Yu Bin, Wu Shanhua, Wang Minghua. K-nearest neighbor model of short-term traffic flow forecast. *Journal of Traffic and Transportation Engineering*. 2012; 12(2): 105-111.
[16] Chakraborty K, Roy I, De P, Das S. Controlling the Filling and Capping Operation of a Bottling Plant using PLC and SCADA. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*. 2015; 3(1): 39-44.