International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.303-312
http://dx.doi.org/10.14257/ijdta.2015.8.6.27

Prediction of Traffic Flow Combination Model Based on Data Mining

[1]Xiaofeng Li and [2]Weiwei Gao

[1,2]*Department of Information Science, Heilongjiang International University, Harbin 150025, China*
[1]*mberse@126.com*,[2]*gvv0451@163.com*

Abstract

It is an important to quickly and accurately forecasting road network traffic flow in intelligent transportation systems, Aiming at the forecasting problem of short-term traffic flow, this paper proposed a traffic flow prediction algorithm, which based on traffic flow sequence partition and neural network model. Firstly, the algorithm divided the traffic flow into different patterns and time sequence by clustering, secondly, described and predicted traffic flow model according to BP neural network. Finally, the experiment shows that based on combined model is much accurate.

Keywords: *Traffic flow, Intelligent transportation system, Data mining, BP neural network*

1. Introduction

Traffic flow prediction is a key part and core content of intelligent transportation system as well as the important basis for transportation information service, traffic control and guidance [1-2]. Forecasting timely and accurately is premise of the intelligent transportation system realizing dynamic traffic management. Crossroads are the key component of transportation network. The size of traffic volume in intersections decides directly the passage capacity of road network, which becomes the bottleneck of road transportation network and plays a significant role in the entire road transportation network [3-4]. To solve the problem of predicting the short-time traffic flow in crossings, it proposes a mining algorithm, which experimentally shows good performance in the real transportation data set [5-6].

According to the time duration of prediction, traffic volume predication can be long-time and short-time predication; as far as the predicted object is concerned, there is crossroad traffic prediction and high-way traffic prediction. Road transportation system is a huge and complicated nonlinear system which has human be involved and is time-varying. The system has higher uncertainties, which may derive from environmental factors like road condition, climate changes *etc.* or emergency situations like traffic accidents, mass gathering *etc.* Those factors bring about certain difficulties to the anticipation of road traffic flow, especially for the short-time prediction. For that, researchers have presented lots of models such as ARIMA model and nonparametric regressive model, which are specifically designed to predict highway and road segment traffic volumes [7-8]. Crossroads are important to the road transportation network. Its transportation is very complex. The traffic flow in each direction in the crossing roads is relevant to not only its own traffic flow and also flow in other direction and the timing plan of traffic signal lights. The traffic flow in crossroads is more volatile than flow series in road sections, particularly the short-time flow series [9]. The volatility occurs not merely because of more accidents taking place in the intersections but also affected by traffic signal lights, specifically variable timing scheme of the signal lights. At present, the acquisition coils for the traffic characteristics of urban intelligent transportation

ISSN: 2005-4270 IJDTA

system in each country are deployed in the crossroads. So the research of short-time traffic flow prediction in such places is one of the important research contents of traffic volume forecasting and of great significance [10-11].

2. Traffic Flow Prediction Algorithm

Regarding the problem of predicting short-time traffic flow in road junctions, it proposes the traffic flow prediction algorithm based on the combined model of traffic flow sequence segmentation and neural network. The algorithm adopts clustering approach to divide the pattern of traffic flow in terms of throughput and time; then with neural network method, it creates models for each traffic flow pattern and makes predictions. The experiment proves that the combined model has better predictive precision to the single neural network model.

2.1. Traffic Flow Sequence Segmentation

Traffic flow data is a form of time sequence. They are segmented according to time sequence data features. It's a hot concern in the time sequence analysis and research field in recent years. The most common method for partitioning time series is piecewise linear description [12-13], *i.e.* applying the linear model to split sequences and make piecewise presentation. Another method is raised based on the probability density function [14]. It utilizes the sufficient statistics of different models to depict subsequence. Here we apply the method based on flow clustering to break up traffic flow sequence. We choose K-Means clustering algorithm as the basic algorithm for sequence segmentation. The basic idea is described as follows:

Step 1: In accordance to the size of values of traffic flow data, select K-Means algorithm as the basic algorithm of clustering segments (set k=3); use Euclidean distance as the function of clustering;

Step 2: As per Step 1, generate cluster C_1, C_2, C_3; then according to $C_1 > C_2 > C_3$, use clustering algorithm to gather C_1 to two clusters, C_2 to three clusters and C_3 to two clusters; but this time by the size of time instead of the traffic flow value, C_1 and C_2 are gathered to generate five clusters;

Step 3: After the above steps, aggregate one-day traffic flow value to seven clusters. Though there's overlapping between clusters, it's possible to discern the order based on the time;

Step 4 : Calculate the centroid and quality of every cluster; the abscissa of centroid is time and the ordinate is mean value of the traffic flow; the quality of clusters is the number of objects in them; order by the size of time in the horizontal axis and mark as $t_1, t_2, t_3, t_4, t_5, t_6, t_7$ and the relative quality as $m_1, m_2, m_3, m_4, m_5, m_6, m_7$.

Setp5. According to the formula 1, calculated $t_1^*, t_2^*, t_3^*, t_4^*, t_5^*, t_6^*, t_7^*$

$$t_1^* = t_1 + \left| \frac{m_1}{m_2 - m_1} \right| \times (t_2 - t_1) \tag{1}$$

2.2. BP Neural Network Prediction Model

Smith *et al*. compared MA, BP neural network and nonparametric regression method in the application of short time traffic flow prediction. They found BP neural network realized higher accuracy rate than traditional Autoregressive Integrated Moving Average (ARMIA) model and was slightly inferior to nonparametric regression method. The neural network method made certain achievements in the research of traffic forecasting.

We selected BP neural network as the fundamental predictive model for short-time traffic flow in crossroads.

BP neural network model has the following characteristics: neurons in each layer connect only with those in neighboring layers; no any connection exists in neurons in the same level; no feedback connection exists in nerve cells in each level. BP neural network is composed of three layers: input layer, hidden layer and output layer. The hidden layer can be omitted according to the requirement of research objective. The initial weight value and threshold value are defined arbitrarily. Learning refers to adjusting gradually the weight value and threshold level to make network's actual output consistent with expected output. The weight coefficient of BP neural network without hidden layer, characteristic of regression coefficient, can be made for regressive application. But it's required to include the hidden layer in BP neural network for classifying and prediction as to increase the degree of approximation. It is shown in Figure 1.

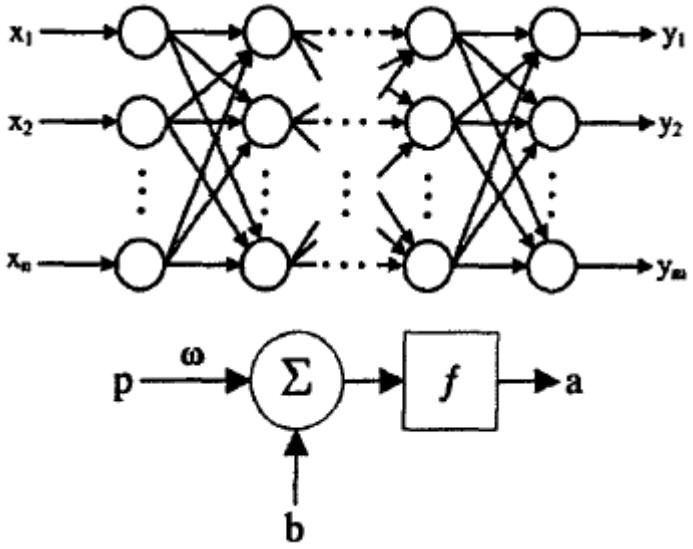

Figure 1. Diagram of Single Neuron Model and the Structure of BP Neural Network

The basic idea of BP algorithm is by processing training sample sets to calculate the output results and the difference between outputs and anticipated results (*i.e.* known category of each sample); then for each training sample, modifying from time to time the weight to minimize the mean square error between network output and actual classes. The error is reversely propagated from input layer to each node. Compute one after another the referential error of each join node and revise the link weight to make the network adapt to the mapping from output error to link weight. The network learning process includes forward propagation and error back propagation.

Forward transmission: calculate input and output of each unit in hidden layer and output layer;

Error reverse propagation: update weight value and offset to reflect the error of network prediction as to spread reversely the error.

2.3. Combined Model of the Traffic Flow Prediction

So far, neural network has limitations, which are listed as follows. Firstly, neural network model is composed of multiple groups of piecewise functions. Hence the prior knowledge between input and output data is hardly integrated into the model, which will affect the precision of recognition. Secondly, in the training process, plentiful original

clean data are required, which presents difficulties to the creation of the model. Thus, BP neural network method will have problems when put for practical engineering applications. In the forecasting of short-time traffic flow, when road network condition and traffic condition change, the well-trained network requires proper updates before being used for the prediction. Meanwhile, no adequate theoretical guidance is available regarding the selection of existing network types and determination of network structure, which are based on experience. Moreover, neural network structure is too complicated. Accuracy rate and approximation ability are relatively higher. But computation and training take a longer time. There are difficulties in the real-time prediction. Yet, too simple network structure can't for sure meet the requirement for precision. Each prediction model has its feasibility. Considering the complexity and particularity of traffic flow in intersections unlike other sites, BP neural network has to improve the accuracy if it's employed solely to predict short-time traffic flow in crossings. Therefore, we should create a multi-model algorithm to increase predictive precision.

The combined model of traffic flow is outcome of the above traffic flow sequence segmentation and BP neural network, whose idea is: on the basis of traffic flow sequence segmentation algorithm, use BP neural network as prediction algorithm to model and foresee the obtained data in given time frame.

3. Experiment Analysis and Results

Use one-week traffic flow data in Turing avenue and southern direction of Cheng-xi road of one city (Figure 2) as the training sample of BP neural network. Then according to the urban traffic monitoring and management system (*i.e*. iCentro-View: The Centro-View system is he city traffic monitoring and management system according to the city traffic of our country current situation of the development of intelligent traffic design) designed as per the basic situation of urban transportation and the development status of intelligent transportation in our country, establish BP neural network for each division section. In the iCentr0View, the coil collects data every 6 minutes, obtaining altogether 2015 samples in one week. We regard traffic values in the first five time units in the prediction direction as input of BP neural network and the traffic values in the next time unit as output. To be specific, there are five input variables (which are predicted historical information, including that in the first five, ten, fifteen, twenty and twenty minutes); the quantity of output variable is one (*i.e*. traffic prediction amount in the following five minutes). BP neural network chooses three-layer neural network structure: input layer, hidden layer, output layer. Output layer has seven units. The prediction results of traffic flow are portrayed in Figure 3.

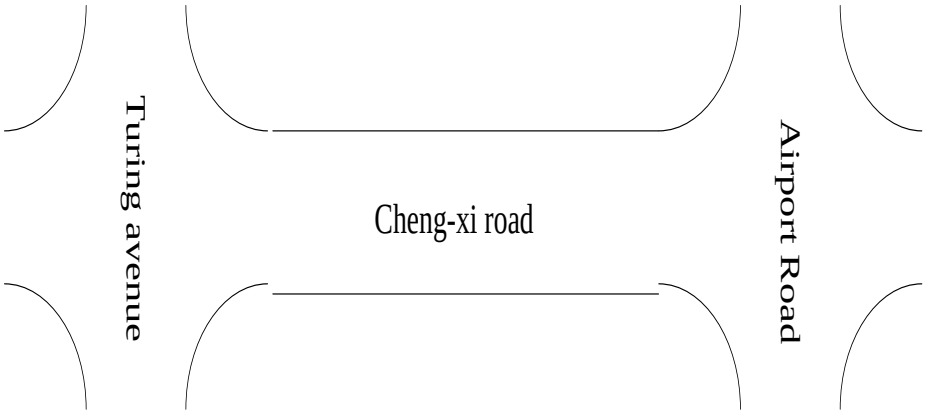

Figure 2. Road Example

In Figure 3, the upper dotted line stands for the prediction error rate with no use of second clustering algorithm, which is 25.48%; the 7-segment function in solid line is the accuracy rate of prediction after the model is created individually for each time frame after the time is segmented, with traffic information in the first 25 minutes as input of the prediction model. Table I illustrates this.

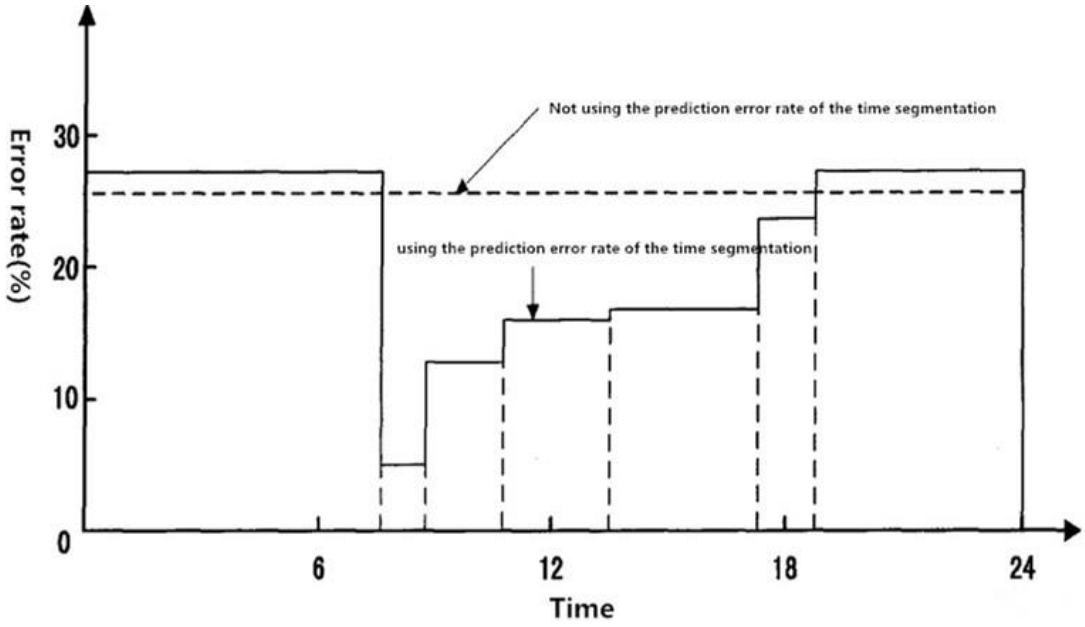

Figure 3. Traffic Flow Prediction Based on Sequence Segmentation

Table 1. Based on Accuracy of Traffic Flow Sequence Segmentation Flow Prediction

Time slot	0:00-7:15	7:15-8:35	8:35-10:30	10:30-13:00
Error rate (%)	24.56	2.14	14.56	18.91
Time slot	13:00-17:00	17:00-18:30	18:30-24:00	
Error rate (%)	14.30	25.10	30.32	

With the combined model of sequence segmentation and BP neural network to predict traffic flow, we find that classifying the mode of traffic flow by two-dimensional clustering algorithm of flux and time conforms to the traffic flow distribution. Through observations, we learn that the traffic flow is distributed in two-peak shapes: morning peak and evening peak. By clustering the size of traffic flow values, we can divide them into several classes. And that we can discern the quantity of flow in the morning peak. The flow includes the maximum information of transportation situations, especially for major urban junctions, traffic signal timing plan is generally not unchangeable. The plan is adjusted automatically as per traffic flow in each direction in the crossroad. The adjustment can be made by fixed time range and timing scheme or adaptively or even done by the control of filter in trunk roads. The change of those statuses can be indirectly reflected by the rate of traffic flow. Besides, for clustering classification in temporal dimension, the data respectively in the morning and evening rush hour can be clearly partitioned.

Experimental results indicate that in the light traffic period, *i.e*. early morning and late night, two prediction methods showed no big differences; but in other time ranges, the traffic flow prediction approach based on the combined model realized very lower error rate, especially in the morning peak, the error rate reduced below 3%. In this way the

precision rate of predicting traffic in rush hours achieved 80%. So the proposed method can satisfy the requirements of traffic control and prediction.

Road traffic flux represents the development or change stage, status and level of the transportation system, which is often expressed with understandable pictorial language or high-level quantitative index like unblocked/congested, normal/abnormal, service level *etc.* Figure 4 sorts the transportation information in road traffic network to some levels from low to high:

(a) Basic data level, referring to data like traffic flow and speed detected by different detection devices;

(b) State feature level, such as traffic status, traffic incident and incident type, traffic jams;

(c) Decision-level information, which is descriptive information about the road network transportation, like service level of road network, impact evaluation of incident and prediction of situational development. The hierarchical structure of traffic information from low to high level can be regarded as the process of data mining or pattern recognition, where low-level information is the foundation to the processing and application of high-level information.

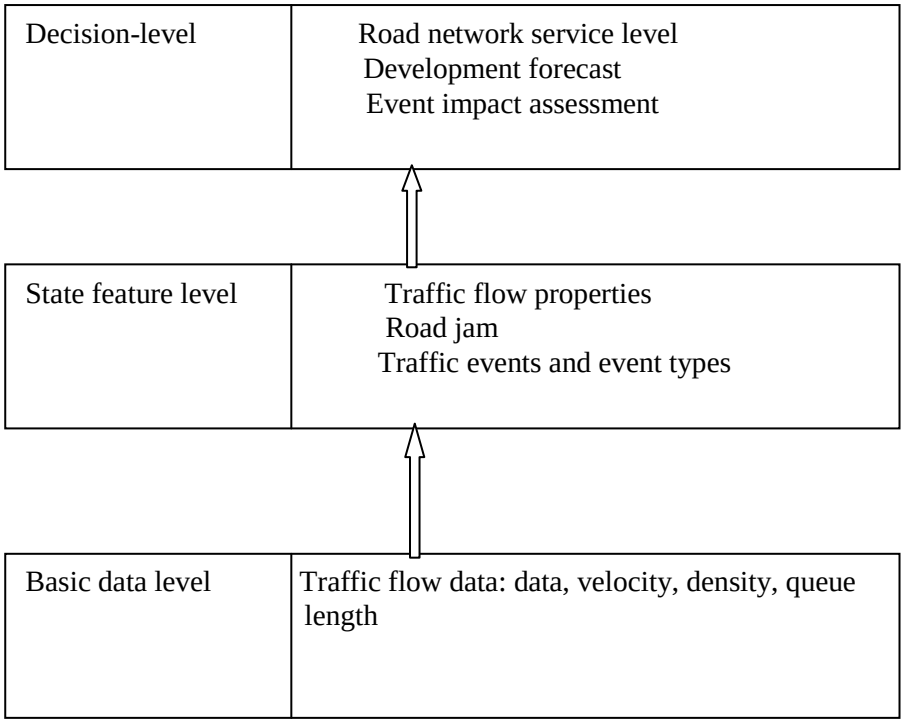

Figure 4. The Traffic Information Level

It's possible to get traffic flow status features by processing traffic data in basic data level. The attribute of traffic flow can be speed, density, occupancy, flow rate, queue length. Single traffic flow attribute can't display the state of ongoing traffic, like the volume, speed. Some papers based on the macro traffic flow model to design the discriminative approach for the traffic status. By focusing on the total traffic volume, static characteristics of roads and motion features of vehicles like speed, journey time, they proposed quantitative index of traffic status recognition. However, the traffic flow model is built as to simplify reasonably discussed topic. If the transportation system is assumed to adapt to some ideal physical model, static index won't reflect the actual running state of road traffic. So the traffic flow recognition method based on fuzzy logic

was raised, which uses self-organizing neural network model to categorize the status of traffic flow. However, regardless the method based on fuzzy logic or self-organizing neural network model, it should utilize the prior knowledge of existing roads' traffic status to perform classification. Such knowledge is subjective and dynamic, acquired very difficultly.

To present well the road traffic state with the traffic flow clustering model, it's required to choose traffic flow parameter which changes apparently with the traffic status like vehicle speed decreasing remarkably when in congestion. The parameter is used as the feature attribute of traffic flow clustering analysis. [15] mentioned the four traffic features which are the most susceptible and have the greatest influence on traffic jams: occupancy ratio, vehicle flow rate, vehicle average speed, time headway. Additionally, the clustering algorithm for clustering analysis affects the quality of the traffic flow clustering model. Currently, for different application purposes, nearly hundred kinds of clustering algorithms were proposed. In the experiment here, we chose one clustering algorithm other than designing a new one. We compare different clustering results as to find out a suitable clustering model for traffic flow prediction.

Choose K-Means clustering algorithm as the fundamental method to make clustering analysis of the traffic flow. Utilize the traffic flow dataset (totally 9410 samples) acquired in Datong road segment of one city in one month for the experiment. We select occupancy rate and traffic volume as the key attribute of clustering analysis, because they are greatly relevant with traffic congestion and more sensitive to the varying of such congestion. Also, they can help collect data of better quality, which is contributive to obtain satisfactory clustering results. The value of k in the K-Means algorithm is determined by experience. We chose empirical value (k= 5, 7 and 9) for testing. The clustering results are listed in Table 2-4 When k=5, the result is ideal, best showing the relationship between traffic volume and occupancy in the congestion period in actual situation.

Table 2. The Cluster Analysis Results of Traffic Flow (k=9)

The cluster label	The number of clusters of objects	The clustering center	
		Flow	occupancy
1	58	17.12	29.14
2	1978	20.12	1.34
3	980	11.25	1.09
4	1756	18.45	1.17
5	2456	20.78	1.36
6	567	34.56	2.14
7	1657	26.78	1.77
8	167	18.45	15.46
9	80	0.12	0.0189

Table2 reveals that the model reflects the change of traffic status: with increasing occupancy ratio, the traffic volume becomes bigger and then tends to go down. Specifically, in the second and fifth cluster, the occupancy rate is higher and flow rate is lower. So based on the knowledge in the transportation field, we can learn that they belong to congestion state; the congestion in the second cluster is much heavier than the fifth; the remaining three clusters are not blocked.

Table 3. The Cluster Analysis Results of Traffic Flow (k=7)

The cluster label	The number of clusters of	The clustering center

	objects	Flow	occupancy
1	2556	19.78	1.45
2	1567	15.05	1.14
3	156	20.56	16.45
4	3089	22.09	1.45
5	85	0.54	0.06
6	88	15.19	29.80
7	1068	30.57	1.98

Table 4. The Cluster Analysis Results of Traffic Flow (k=5)

The cluster label	The number of clusters of objects	The clustering center	
		Flow	occupancy
1	1787	14.14	1.09
2	89	18.16	30.87
3	2567	28.79	1.89
4	4678	21.09	1.56
5	156	18.79	14.54

4. Conclusion

The paper discussed the problem of predicting short-time traffic passing in crossroads. To solve it, it proposed the combined model prediction approach based on second clustering traffic flow sequence segment and BP neural network. The experiment confirmed that the new solution reached higher precision rate and can forecast accurately the traffic discharge in the following five minutes in each direction of the intersection.

Acknowledgements

This work is supported by The Education Department of Heilongjiang province science and technology research project, Under Grant No. 12543076

References

[1] L. Lina, “Research on city road traffic flow short-time prediction”, Beijing University of Posts and Telecommunications, **(2010)**.
[2] F. Steaming, “Research on city road traffic flow short-time prediction”, Beijing Jiaotong University, **(2012)**.
[3] L. Jiaqi, “Development and application of automatic detection system for road traffic flow”, Hunan University, **(2013)**.
[4] T. Qian, “Research on dynamic traffic assignment model of urban road network emergencies”, Jilin University, **(2013)**.
[5] S. Yixuan, “Research”, Analysis of road traffic accident based on data mining of Beijing Jiaotong University, **(2014)**.
[6] C. Juan, Z. Yingchun and S. Bohong, “The road traffic network flow entropy optimization based on the vulnerability”, Journal of Southwestern Normal University (Natural Science Edition), vol. 7, **(2014)**, pp. 30-35.
[7] X. Xu and C. Jinbo, “Effect of road network traffic flow distribution of structural reliability”, the highway traffic science and technology, vol. 7, **(2014)**, pp. 121-128.
[8] S. X. Liang, “City road traffic state evaluation and prediction method and application research”, Beijing Jiaotong University, **(2013)**.
[9] X. Bin, “Traffic flow monitoring and implementation of city road based on video analysis”, Nanjing University of Posts and Telecommunications, **(2013)**.
[10] W. Yong, “Research and analysis”, the city area traffic signal control and traffic state of Zhejiang University, **(2013)**.
[11] W. Mofi, “The city intelligent traffic system, traffic flow and the key technology of collaborative optimization induction”, Central South University, **(2013)**.

[12] E. K. Chu, "An online algorithm for segmenting time series", Proceeding of the IEEE international conference on data mining, **(2001)**, pp. 289-296.
[13] E. K. Kasetty, "On the Need for Time Series Data Mining benchmarks: A survey and empirical demonstration", in proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, **(2002)**, pp. 23-46.
[14] J, Lu and L. Cao, "Congestion evaluation from traffic flow information based on fuzzy logic", IEEE, **(2003)**, pp. 345-348.
[15] G. Tan, "Traffic flow prediction based on generalized neural network", 2004 IEEE intelligent transportation system conference .Washington .D.C, USA, **(2004)**, pp. 45-48.

Author

Xiaofeng Li, He is an advanced member of china computer federation and he is an associate professor in Heilongjiang International University. His research interest includes data mining and intelligent algorithm.