Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 96 (2013) 653 – 662

Procedia
Social and Behavioral Sciences

13th COTA International Conference of Transportation Professionals (CICTP 2013)

An Improved *K*-nearest Neighbor Model for Short-term Traffic Flow Prediction

Lun Zhang, Qiuchen Liu, Wenchen Yang, Nai Wei*, Decun Dong

School of Transportation Engineering, Tongji University, Shanghai, 201804, China

Abstract

In order to accurately predict the short-term traffic flow, this paper presents a k-nearest neighbor (KNN) model. Short-term urban expressway flow prediction system based on k-NN is established in three aspects: the historical database, the search mechanism and algorithm parameters, and the predication plan. At first, preprocess the original data and then standardized the effective data in order to avoid the magnitude difference of the sample data and improve the prediction accuracy. At last, a short-term traffic prediction based on k-NN nonparametric regression model is developed in the Matlab platform. Utilizing the Shanghai urban expressway section measured traffic flow data, the comparison of average and weighted k-NN nonparametric regression model is discussed and the reliability of the predicting result is analyzed. Results show that the accuracy of the proposed method is over 90 percent and it also rereads that the feasibility of the methods is used in short-term traffic flow prediction.

*Keywords:*Prediction; Short-term Traffic flow; Nonparametric Regression Model; *k*-NN; Urban Expressway.

1. Introduction

Real-time, accurate traffic flow prediction is the key to traffic control, traffic induction, and providing real-time traffic information. Effectiveness of traffic information and the accuracy of the detection of abnormal events and so intelligent transportation systems, traffic signal control of real-time traffic information released traffic predict studies are based on the short-term traffic flow predict (less than 5min). The short-term traffic flow predict uses the existing traffic flow data at time t to estimate traffic flow at the next time t + Δt (Smith, Williams, and Keith, 2002).

* Nai Wei. Tel.: +86 18621908884;
E-mail address: niwei.tongji@gmail.com

doi:10.1016/j.sbspro.2013.08.076

Methods of short-term traffic flow prediction can be divided into two categories at home: One is based on traditional mathematics and physics models, including the historical average model (Keith, Scherer, and Smith, 2000), time series models (Park and Rilett, 1998), Kalman filtering model (Nihan and Holmesland,1980) and exponential smoothing model etc.; the other is the mathematical model of the predicting methods, neural networks, nonparametric regression, and support vector model (Dasarathy, 1991) are included.

Because traffic flow is uncertain, nonlinear and complex, it is difficult to predict the traffic flow effectively and accurately by the predict method based on traditional mathematics and physics models (Armstrong, 2006). In contrast, the non-mathematical model prediction methods, such as neural networks, support vector machines, etc., are easy to build and have more accurate prediction. But these methods need to do time-consuming adjustment and the portability is poor. From the point of view that above prediction methods has shortcomings; K-nearest neighbor nonparametric regression methods are put forward to predict short-term traffic flow. Extensive prediction results have demonstrated the accurate and portable potential of the proposed prediction method for short-term traffic flow.

2. Basic theory on traffic flow prediction

2.1 Characteristics of urban Expressway traffic flow

Urban expressway is the highest level of road in the city, which ensures that drivers can travel quickly and continuously. Characteristics of traffic flow are as follows:

(1) Periodicity. Traffic flow shows cyclical changes.
(2) Randomness. Traffic system is a uncertain system affected by multiple factors, such as travel behavior, weather, accidents, etc. Thus, there are some differences between the existing data and historical data.

2.2 K-nearest neighbor nonparametric regression algorithm

K-nearest neighbor nonparametric regression method is a broad applied algorithm, which has nonparametric, small error ratio and good error distribution (Yiannis and Poulicos, 1857).

The basic process of k-nearest neighbor prediction modal is shown in Fig. 1. First, build a representative historical database with large capacity; second, set the model elements, including the state vector value of k and prediction algorithm. The state vector and value of k constitute a model's search mechanism. Finally, according to the observed values of the input and search mechanism, a close neighbor matching the current real-time observation data from the history database are picked up to predict the traffic flow at the next time (Smith and Demetsky, 1994).

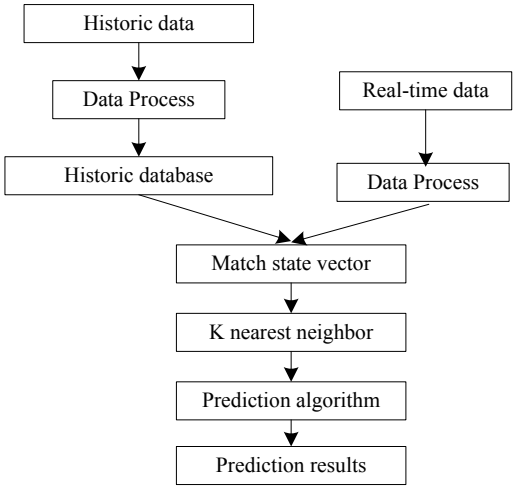


Fig. 1 Process of k-nearest neighbour nonparametric regression algorithm

3. K-nearest neighbor prediction method

3.1 Historical database

3.1.1Data preprocessing

Prediction results of *k* nearest neighbor nonparametric regression depend directly on the quality of the sample database. In order to get a streamlined and effective historical database, the collected basic data (traffic flow, traffic density, traffic speed) of urban expressway needs to be pre-processed. The data preprocessing method is as follows:

(1) Check duplicate data
(2) Reasonable range of traffic flow: Volume, Speed and Occupy.
1) Reasonable range of volume:

$$0 \le Volume \le m_c \times CAP \times T / 60 \tag{1}$$

Here, CAP is road capacity (veh/h); T is cycle of data collection(min)； mc is correction factor
2) Reasonable range of average speed:

$$0 \le Speed \le m_v \times Speed_1 \tag{2}$$

Here, *Speed* is the limit speed of the road, m_v is a correction factor, and here m_v is a random number within [1.3~1.5].

3) Reasonable range of occupy:

$$0 \le Occupy \le 100\% \tag{3}$$

The abnormal data identified according to the above threshold method will be removed, and the adjacent data replace the abnormal data using numerical interpolation.

(3) Test data consistency

A record with share zero, while the volume of traffic and the speed is not zero will be abandoned.

(4) Data loss

If repaired data is unable to be verified correct, the data will not be repaired.

3.1.2 Data standardization

After data preprocessing, traffic flow data *Volume* needs normalization, which can avoids the magnitude difference of the sample data and prediction deviation ,so that the prediction accuracy can be improved.

$$Volume^* = \frac{Volume - Volume_{\max}}{Volume_{\max} - Volume_{\min}} \tag{4}$$

$$Volume_{\max} = \frac{1}{n}\sum_{j=1}^{n} Volume_{\max j} \tag{5}$$

$$Volume_{\min} = \frac{1}{n}\sum_{j=1}^{n} Volume_{\min j} \tag{6}$$

Here, $Volume^*$ is normalization data, ranging between [0, 1]; $Volume_{\max j}$ is the maximum volume on day j; $Volume_{\min j}$ is the minimum volume on day j and n is the number of sample days.

3.2 Search mechanism and algorithm parameters

3.2.1 Prediction time interval

Cycle of traffic control is 2.5~3min and cycle of traffic induction is generally 5 min. Therefore, how to accurately predict traffic flow within 5min is the key to traffic control and traffic induction. This study selects 5 min as prediction time interval.

3.2.2 State vector and hysteresis values q

(1) State vector

State vector is a standard for comparing the observed data and the historical database and state vector will eventually affect the value of near neighbor, which is directly related to the prediction accuracy. Occupancy

depends on the flow and speed, so in speed stability models more single case, the trend of the share of traffic should be broadly similar.

Within a certain range, the low correlation between the traffic flow and speed presents the change trend of occupancy and flow is substantially same. Therefore, traffic flow is selected as the state vector.

(2) Hysteresis value q

The continuous-time hysteresis value *q* is selected as a state vector to match the historical data to avoid resulting in excessive similar values when comparing the current observation data and the historical database. Flow of continuous time *t*, *t-1*,… *t-q* will be state vector X_t.

$$X_t = \{V_t, V_{t-1}, ...V_{t-q}\} \tag{7}$$

3.2.3 Nearest neighbor K

K-nearest neighbor method is non-parametric regression method which searches for the *k* optimal nearest neighbor and predicts traffic flow at the next time.

Too large or too small value of *k* will affect the prediction accuracy. The value of *k* is set between 5 and 30 in preliminary experiments. Finally, experimental results will show whether the value of *k* is reasonable.

3.2.4 Prediction algorithm

After the above matching mechanism, Assuming that the *k* nearest neighbors are found in the history database, distance between the actual data and the *k* nearest neighbor is $q_i (i = 1, ..., k)$ and then the flow at the next moment is $Volume_{hi}(t+1)\ (i = 1, ..., k)$. The widely used prediction algorithms are algorithms with non-weighted and algorithms with weighted. Algorithms with non-weighted is as follows (Guo Jianhua, 2005):

$$Volume_{t+1} = \frac{1}{k}\sum_{i=1}^{k} Volume_{h,i,t+1} \tag{8}$$

Algorithms with weighted is as follows:

$$Volume_{t+1} = \sum_{i=1}^{k} a_i Volume_{h,i,t+1} \tag{9}$$

$$a_i = \frac{q_i^{-1}}{\sum_{i=1}^{k} q_i^{-1}} \tag{10}$$

This paper uses these two algorithms to predict traffic flow.

3.2.5 Reduction of prediction data

The prediction results of k-nearest neighbor prediction mechanism are standardization data, which needs reduction to modulus data. The conversion formula between standardization data and modulus data is shown below. Data standardization avoids the magnitude difference of the sample data and prediction deviation, so that the prediction accuracy can be improved.

$$Volume' = Volume*' \times (Volume_{\max} - Volume_{\min}) + Volume_{\min} \tag{11}$$

Here, *Volume*'* is standardized prediction data; *Volume'* is the prediction data.

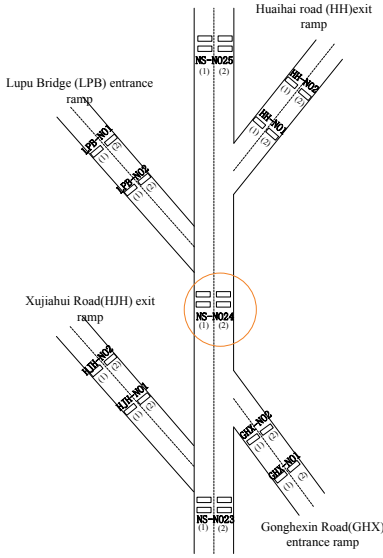
4. Case study

4.1 Data source

The basic data source comes from traffic information detection system of Traffic Information Center of Shanghai, which covers traffic data from inter-city highways**,** expressway network and the main roads of Shanghai. In this study, the North-South Elevated Road of Shanghai with 21 ramps is selected as the research case in this study. As shown in Fig. 2(a), the North-South Elevated Road of Shanghai gets through the city center of Shanghai from north to south and connects the Central Expressway, outer ring expressway and inner ring expressway of Shanghai.

As shown in Fig. 2(b), traffic flow data comes from the deployed loops NS-NO24, (1), (2) in North-South Elevated Road of Shanghai, from Gonghexin Road to Huaihai Road. The time interval of the collected data is 20s. Data collection time is from October 1, 2012 to November 19, 2012. And traffic flow data from October 1, 2012 to November 17, 2012 is selected as the historic database and the data collected from November 18, 2012 to November 19, 2012 is used as test data.

(a). North-South Elevated Road(NS), Shanghai



(b). Diagram of deployed loops between GHX and HH ramp

Fig.2 Diagram of loops on North-South Elevated Road, Shanghai

4.2 Evaluation

In order to evaluate the performance of the proposed model, *MAD* (Mean Absolute Difference) and *MAPE* (Mean Absolute Percent Error) are introduced (Fan Wang, 2010).

(1) *MAD* (Mean Absolute Difference)

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|Volume_i^{'} - Volume_i\right| \tag{12}$$

The larger the value of *MAD*, the greater the prediction error; In contrary, the more accurate the prediction accuracy.

(2) *MAPE* (Mean Absolute Percent Error)

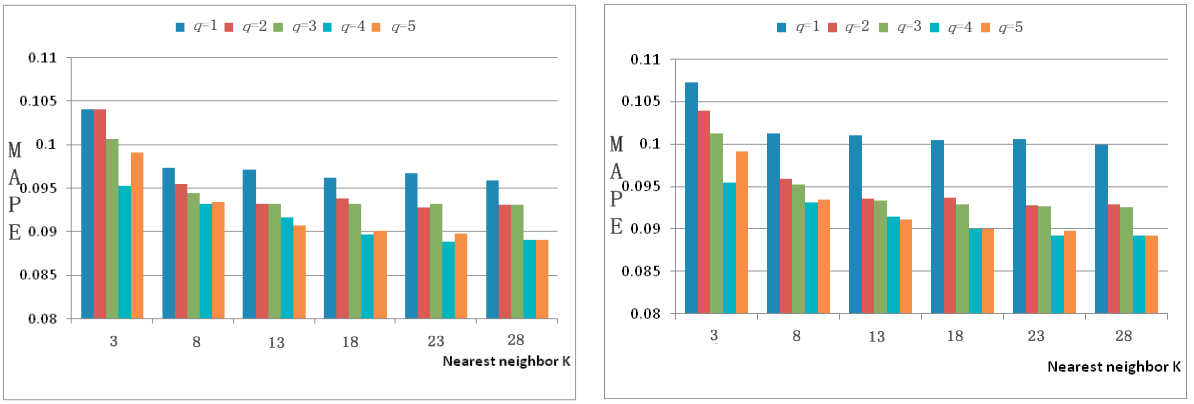
$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Volume_i^{'} - Volume_i}{Volume_i}\right| \tag{13}$$

The larger the value of *MAPE*, the greater the prediction error; In contrary, the more accurate the prediction accuracy.

4.3 Results

4.3.1 Determination of value of k and value of q

This paper selects *MAPE* as measurement standard and selects hysteresis value from 3 to 28. As shown in Fig.3, results of the equal weight and exponent weight algorithms are compared.



(a). Algorithm with non-weighted (b). Algorithm with weighted

Fig. 3 Effects of value of k and *q* on prediction accuracy

This paper preliminary makes conclusions that K-NN nonparametric regression algorithm gets the best accuracy when *k* takes about 18, and hysteresis value *q* takes about 4.

4.3.2 Result Analysis

This study selected a set of predictive value ($k = 18$, $q = 4$) with the least MAPE and analyzed the prediction results from loop NS-NO24, (1), (2) in North-South Elevated Road of Shanghai, from Gonghexin Road to Huaihai Road from November 18, 2012 to November 19,2012. The observed traffic flow data and the predicting traffic flow data are shown in Fig. 4 and Fig. 5.

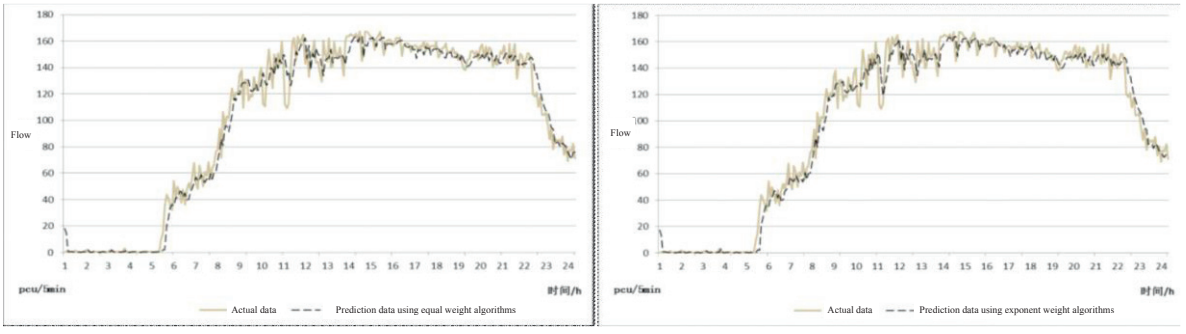


Fig. 4 Prediction results of November 18, 2012

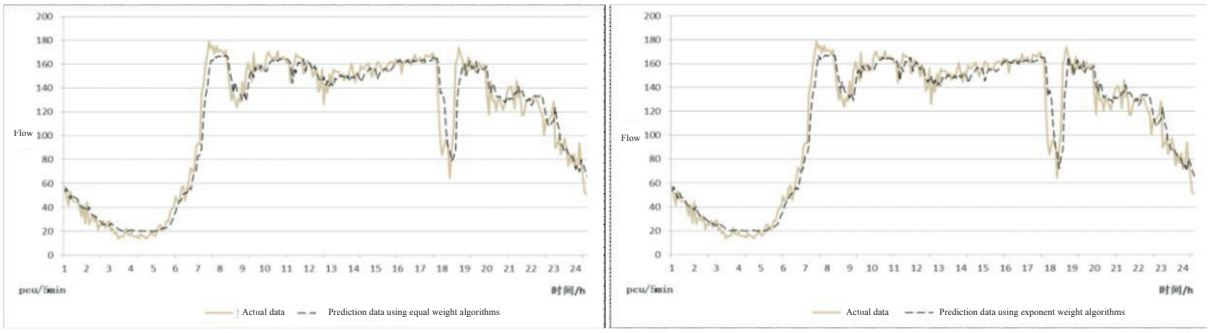
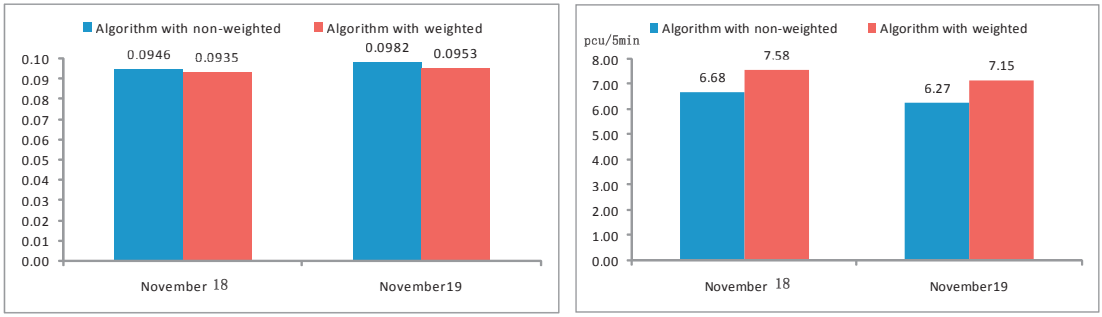


Fig. 5 Prediction results of November 19, 2012

As shown in Fig. 6, this paper proposes urban expressway traffic flow prediction method based on k-NN nonparametric regression of which MAPE is less than 10 percent. Results show that the proposed predicting method can meet the real-time and accurate requirements of traffic flow.



(a). Histogram of MAPE (b). Histogram of MAD

Fig. 6 Comparation of prediction results of algorithms with non-weighted and weighted

5. Conclusion

This paper presents a k-nearest neighbor (KNN) model. First, traffic flow demonstrates uncertain, nonlinear and complex characteristics. Performance indices prove that the proposed prediction method can be used to short-term predict of road traffic flow traffic flow. The short-term urban expressway flow predicting system based on k-NN is established in three aspects: the historical database, the search mechanism and algorithm parameters, and the predication plan. Firstly, preprocess the original data and then standardized the effective data in order to avoid the magnitude difference of the sample data and improve the prediction accuracy. At last, a short-term traffic prediction based on k-NN nonparametric regression model is developed in the Matlab platform.

Numerical experiments results show that the accuracy of the urban freeway traffic flow prediction method based on k-NN nonparametric regression is over 90 percent and it also rereads that the feasibility of the methods is used in short-term traffic flow prediction.

Acknowledgment

The work of this paper is supported by National Science Foundation of China (Project No. 50408034) and Shanghai Educational Foundation for Innovation (Project No. 11ZZ27).

References

B. L .Smith, B. M. Williams, O. R .Keith (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C*, 2002(10):303.

Keith, R.W. Scherer, T., Smith, B. L. (2000). Traffic flow forecasting using approximate nearest neighbor nonparametric regression. The National ITS Implementation Research Center U.S. DOT University Transportation Center.

Park,L.& Rilett,R.. (1998). Forecasting multiple-period freeway link travel times using modular neural networks. *Transportation Research Record, 1998,* 1617:163-170.

Nihan, L. & Holmesland, O. (1980). Use of the box and Jenkins time series technique in traffic forecasting. *Transportation Research Record 1980,* 9(2):125-143.

B.V. Dasarathy (1991). Nearest Neighbor Forms: NN Patten Classification Techniques. Los Agapitos IEEE Computer Society Press, 1991.

Yiannis, K. & Poulicos, P. (1857). Forecasting traffic flow conditions in an urban Network-comparison of multivariate and univariate approaches. *Transportation Research Record 1857,* 2003:74-84.

Smith, B. L. & Demetsky, J. (1994). Short-term traffic flow prediction: neural network approach. *Transportation Research Record 1453*, 1994:98-104.

Armstrong, J.S. (2006). Findings from evidence-based forecasting: methods for reducing forecast error. *International Journal of Forecasting 2006*, 22:583-598.

Anthony, S. & Matthew, K.G. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research part C,* 2003, 11(2):121-135.

Guo Jianhua (2005). Adaptive estimation and prediction of univariate vehicular Traffic condition series [D]. North Carolina State University.

Fan Wang (2010). Research on Methods of Traffic Flow Forecasting Based on SVM [D]. Dalian University of Technology.