

Key Factors of K-nearest Neighbors Nonparametric Regression in Short-time Traffic Flow Forecasting

Jing-ting Zhong, Shuai Ling*

College of Management and Economics, Tianjin University, Tianjin, China
(ls5209@163.com)

Abstract - Short-term traffic flow prediction plays an important role in route guidance and traffic management. K-NN is considered as one of the most important methods in short-term traffic forecasting, but some disadvantages limit the widespread application. In this paper, we use four tests to find the key factors of the K-NN method, which will give inspires to the further research to improve the method.

Keywords - K-nearest neighbors, key factors, short-term traffic flow forecasting

I. INTRODUCTION

Shore-term traffic flow forecasting has played an important role in the intelligent transportation system (ITS). Traffic management departments can use it to design strategies of traffic control and traffic guidance, while travelers use it to make route choice. Under the above benefits, short-term traffic flow forecasting has become an attractive field both in traffic science and traffic engineering.

By definition, short-term traffic forecasting is the process to predict key traffic parameters such as speed, flow, occupancy, or travel time with a forecasting horizon typically ranging from 5 to 30 minutes at specific locations.

There are generally two kinds of approaches for short-term forecasting: parametric approach and non-parametric approach. The parametric approach assumes that there is an explicit forecasting mathematic by a set of parameters. Historical data is then used to find out a group of parameters which can minimum the forecasting error (gained by the historical data). Afterwards, the model can be used in the real-time forecasting.

Contrarily, instead of finding the explicit mathematic form the relationship between inputs and outputs, nonparametric approaches are data-driven and allow data to speak for itself^[1] (Bosq 1996). The most popular nonparametric approaches includes nonparametric regression (NPR).

The NPR model is a data-driven approach. Instead of trying to compress all training data into a set of mathematical specifications (parametric approaches) or a certain network (ANN) through modeling process, it retains all historical observation and searches for the most similar case of the current state, based on which forecasting is then made. Oswald, Scherer et al. (2000) investigated the practical use of NPR model, and discussed some problem likely to encounter in the real world usage^[2]. (Clark 2003) studies the multi-variant NPR forecasting, as well as the influence of neighbor size and the transferability of database, which are valuable for

the practical use of NPR model^[3]. (Chang, Zhang et al. 2011) made three improvements, including the data organization and the search mechanism, for faster calculation and higher accuracy^[4]. Other researches considered additional information in NRP forecasting, such as historical traffic state and traffic condition information, and stated these helps to reduce forecasting error^[5,6] (Abdulhai, Porwal et al. 2002; Gong and Wang 2002).

With the continuous deepening researches on K-NN, they become increasingly mature in short-time traffic flow forecasting and will be applied in intelligent transportation system^[7,8] (Friedman, Bentley et al. 1977; Van Der Voort, Dougherty et al. 1996). For example, with the help of nonparametric forecasting models, a real-time, on-line, self-learning forecast system can be implemented^[9] (Zhu and Yeh 1998).

In practical applications, Although K-NN has many advantages, the application of nonparametric regression methods still have to pay attention to that the K-NN method involve massive data and calculations.

II. K-NEAREST NEIGHBORS NONPARAMETRIC REGRESSION

In fact, nonparametric regression is based on pattern matching and data mining. Suppose the short-time traffic flow forecasting has to predict the traffic flow of a section at next time epoch, the corresponding influencing factors ($f_1 \sim f_n$) have to be explored firstly^[10-12] (Bentley 1975; Bentley and Friedman 1979; Bentley 1990). These influencing factors may include traffic flows of the section and upstream section at previous time epoch, weather, road conditions, etc. In this paper, $f_1 \sim f_n$ are taken as the state components of the system, which compose the state vector of the system (f_1, \dots, f_n). The traffic flow at the forecasting time epoch (q) is called as decision attribute, which is determined by (f_1, \dots, f_n). In other words, the current state vector ($F = (f_1, \dots, f_n)$) determines q at the forecasting time epoch. In this paper, F and q composed a pattern vector ($P = \{(f_1, f_2, \dots, f_n) | q\}$).

Search k nearest neighbor (KNN) of the forecasting state (F_{pre}) in the historical pattern according to equation (1) and then predict by using q of KNN.

III. EXPERIMENTAL DATA

The Data used in this research is collected from website of University of Minnesota Duluth (<http://www.d.umn.edu/tdrl/traffic/>). Real-time traffic data have been collected since 1997 from over 4,000 double inductive loop detectors located around the Twin Cities Metro freeways. The road used for experiment is a section of I-35E South, intersecting with TH110 and Wagon Wheel trail. The network structure and detectors layout is shown in Fig.1.

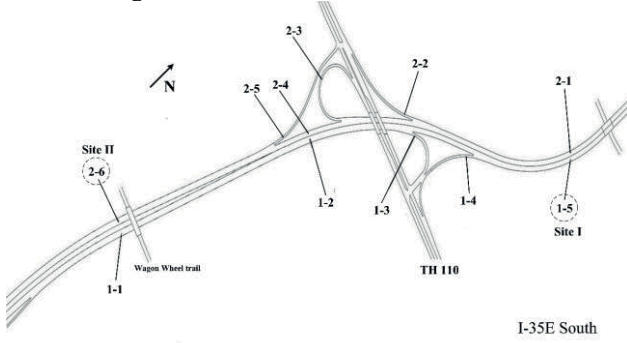


Fig.1. Twin Cities Metro freeways

Traffic data come from University of Minnesota Duluth (<http://www.d.umn.edu/tdrl/index.htm>). Based on the traffic management center of Minnesota, the Traffic Data Research Laboratory of University of Minnesota Duluth can provide all traffic data on expressways surrounding the University of Minnesota-Twin Cities. These data were collected continuously by more than 4,000 loop detectors around the whole year and compressed in exclusive format, enabling to output traffic flow and occupancy recorded by desired detector during the desired period.

A. Pattern composition

In this paper, data collected during 2012.03.01-2012.04.31 were used. We have recoded all the detectors in the map. Traffic flow at detector 1-5 and detector 2-6 are selected for experiments. For convenience, in the remaining of the paper we name them as site1 and site2 respectively.

With respect to the traffic flow forecasting at detector 1-5 and detector 2-6, the test adopted two compositions of state vector of modes: ① simple pattern: For data that is time series in nature, a state vector defines each record with a measurement at time ^[13,14] (Smith and Demetsky 1997; Smith, Williams et al. 2002); ② complicated pattern: both historical flow on time series of the forecasting section and flow of related upstream section are taken into account ^[6,15] (Abdulhai, Porwal et al. 2002; Li, Li et al. 2013).

Time-series pattern and temporal-spatial pattern under different state vectors were presented in the following.

For detector 1-5:

$$P_{Time-series}^{15} = \{f_{15}(t), f_{15}(t-1), f_{15}(t-2) | f_{15}(t+1)\} \quad (1)$$

$$P_{Temporal-spatial}^{15} = \{f_{15}(t), f_{15}(t-1), f_{15}(t-2), f_{12}(t), f_{12}(t-1), f_{14}(t), f_{14}(t-1) | f_{15}(t+1)\} \quad (2)$$

For detector 2-6:

$$P_{Time-series}^{26} = \{f_{26}(t), f_{26}(t-1), f_{26}(t-2) | f_{26}(t+1)\} \quad (3)$$

$$P_{Temporal-spatial}^{26} = \{f_{26}(t), f_{26}(t-1), f_{26}(t-2), f_{24}(t), f_{24}(t-1), f_{25}(t), f_{25}(t-1) | f_{26}(t+1)\} \quad (4)$$

where $f_{ij}(t)$ is the flow detected by detector i j at t and $f_{ij}(t-1)$ is the flow detected by detector i j at $t-1$.

In the KNN system design, nearest neighbors within a ring (centered at the forecasting pattern and $r = 50$) are searched firstly by combining fixed radius and fixed KNN search strategy. If k^* patterns were searched, then the number of nearest neighbors (k) can be determined by the following equation:

$$k = \begin{cases} k^* & k^* < 20 \\ 20 & k^* \geq 20 \end{cases} \quad (5)$$

B. Evaluation index of test results

RMSE is the common evaluation index of test results:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2} \quad (6)$$

where x_i is true value, \hat{x}_i is predicted value and

N is number of patterns. RMSE is affected by the value range of experimental subject and couldn't reflect the absolute error. Therefore, this experiment used the mean absolute error (MAE) for evaluating the test results. MAE is dimensionless and insensitive to the value range of experimental subject.

$$MAE = \frac{1}{N} \sum_{i=1}^N \frac{|x_i - \hat{x}_i|}{x_i} \quad (7)$$

IV. TEST DESIGN

A. Test1

The original traffic flow data collected during 2012.03.01-2012.04.30 were processed firstly to generate corresponding patterns. In the original state vector, states deviated significantly from the group are called as outliers. Although patterns including these outliers are real, they are against the test and deleted by using DBScan algorithm.

Density-Based Spatial Clustering of Applications with Noise (DBScan) is a representative clustering algorithm based on density. It doesn't need to know number of clusters in advance as it can identify all forms of clusters. More importantly, it can recognize outliers (red points in Fig.2).

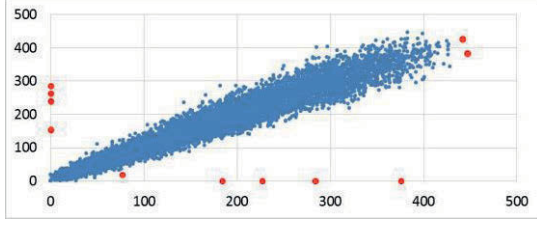


Fig.2. Two-dimensional projection of original state vector distribution (red points are outliers)

After the processing, the candidate data set (C) contains 15,500 data. Divide C in to 100 subsets randomly(C_1, C_2, \dots, C_{100})

$$C_1 \cup C_2 \cup \dots \cup C_{100} = C, C_i \cap C_j = \emptyset \quad i \neq j \quad i, j \in 1, 2, \dots, 100).$$

Both KNN and ANN predicted the traffic flow on 2012.04.31 for 100 times and added a new data subset in the database after every prediction. Next, the data in the new database were used for KNN prediction and ANN training.

Fig.3 shows spatial distribution of state vectors with different data density.

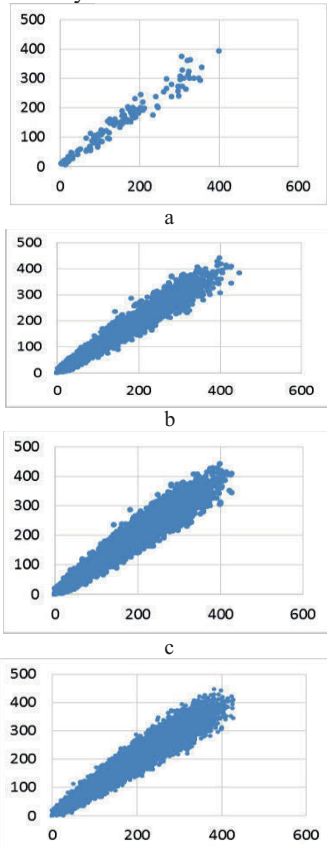


Fig.3. a 1% data density, b 25% data density, c 50% data density, d 100% data density

Test results were listed as Fig.4:

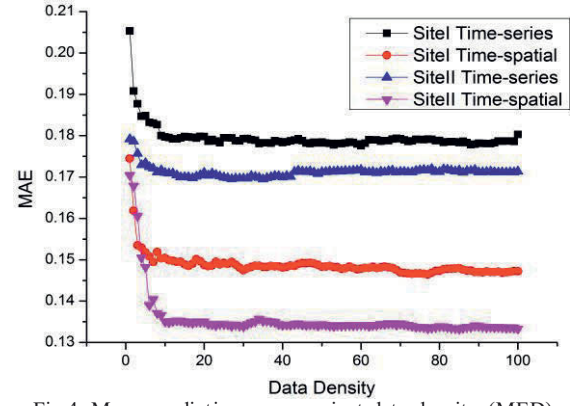


Fig.4. Mean prediction error against data density (MED)

B. Test2

Test design: Divide P into three independent subsets (P_A, P_B, P_C). Fig.5 is the two-dimensional projection of their state space.

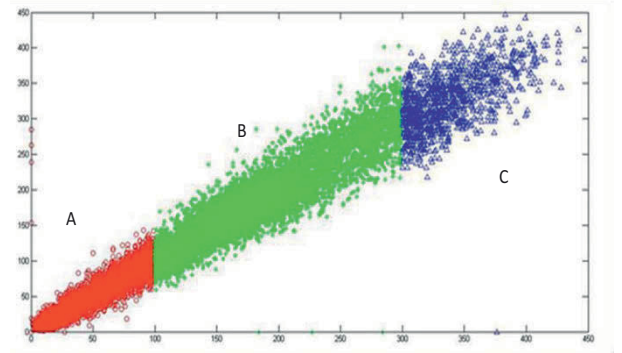


Fig.5. State space division

Decrease data in B at a rate of 1/15 and take the new dataset as the database of nonparametric regression forecasting system and training data of ANN system. Meanwhile, A and C are taken as the test set for prediction.

Test results were listed as Fig.6:

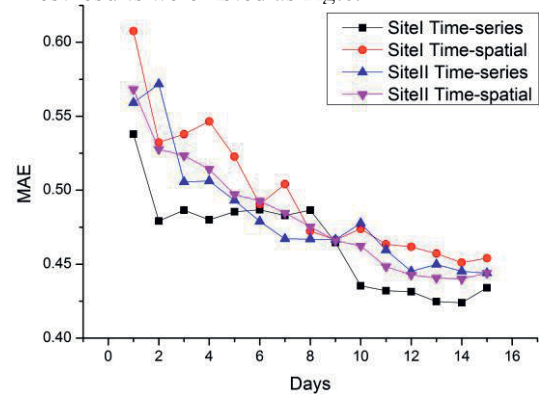


Fig.6. Mean prediction error against data density

C. Inspirations from tests

Viewed from the prediction nature, we can find that KNN system predicate traffic flow by averaging the neighboring patterns of the forecasting pattern. Longer running history of database will bring more similar neighboring patterns of the forecasting pattern.

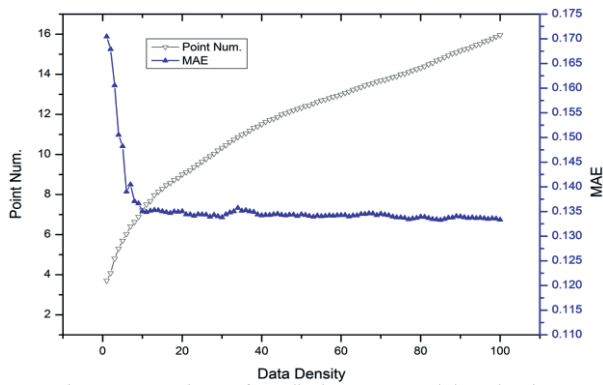


Fig.7. Comparison of prediction error and data density

Fig.7 shows the prediction results of $P_{\text{Temporal-spatial}}$ of KNN system at SiteII with random increasing. Point Num. represents the number of nearest neighbors searched by KNN system within the ring of $r = 45$ centered at the forecasting point and MAE represents the mean prediction error. When data density reached to the data density at turning point, MAE tends to be stable. At this moment, 7 nearest neighbors in average are searched within the ring. Subsequently, excessive nearest neighbors searched within the ring make no contributions to the prediction error reduction.

V. CONCLUSIONS

The short-time traffic flow forecasting performances of KNN and ANN methods are analyzed through tests, finding that:

1. The prediction accuracy of KNN system presents no linear improvement with the increasing of data size in the database. When data size increases to a certain level, the prediction accuracy of KNN will remain basically same.

2. The prediction accuracy of KNN is closely related with the state vector. More information contained in state vector will make it more representative and thereby brings higher prediction accuracy.

Additionally, this paper also explored some ways to further improve the prediction accuracy of ANN and KNN, such as optimizing composition of state vector, deleting data in high-data-density region, using representative data and using time-series data increasing.

REFERENCES

- [1] Bosq, D. (1996). "Nonparametric statistics for stochastic processes." Lecture Notes in Statist.
- [2] Oswald, R. K., W. T. Scherer, et al. (2000). "Traffic flow forecasting using approximate nearest neighbor nonparametric regression." Final project of ITS Center project: Traffic forecasting: non-parametric regressions.
- [3] Clark, S. (2003). "Traffic prediction using multivariate nonparametric regression." Journal of transportation engineering 129(2): 161-168.
- [4] Chang, G., Y. Zhang, et al. (2011). A summary of short-term traffic flow forecasting methods. 11th International

Conference of Chinese Transportation Professionals (ICCTP 2011).

- [5] Gong, X. and F. Wang (2002). Three improvements on knn-npr for traffic flow forecasting. Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on, IEEE.
- [6] Abdulhai, B., H. Porwal, et al. (2002). "Short-term traffic flow prediction using neuro-genetic algorithms." ITS Journal-Intelligent Transportation Systems Journal 7(1): 3-41.
- [7] Friedman, J. H., J. L. Bentley, et al. (1977). "An algorithm for finding best matches in logarithmic expected time." ACM Transactions on Mathematical Software (TOMS) 3(3): 209-226.
- [8] Van Der Voort, M., M. Dougherty, et al. (1996). "Combining Kohonen maps with ARIMA time series models to forecast traffic flow." Transportation Research Part C: Emerging Technologies 4(5): 307-318.
- [9] Zhu, J. and A. G.-O. Yeh (1998). "A self-learning short-term traffic forecasting system." Horizons 72: 13.32.
- [10] Bentley, J. L. (1975). "Multidimensional binary search trees used for associative searching." Communications of the ACM 18(9): 509-517.
- [11] Bentley, J. L. (1990). K-d trees for semidynamic point sets. Proceedings of the sixth annual symposium on Computational geometry, ACM.
- [12] Bentley, J. L. and J. H. Friedman (1979). "Data structures for range searching." ACM Computing Surveys (CSUR) 11(4): 397-409.
- [13] Smith, B. L. and M. J. Demetsky (1997). "Traffic flow forecasting: comparison of modeling approaches." Journal of transportation engineering 123(4): 261-266.
- [14] Smith, B. L., B. M. Williams, et al. (2002). "Comparison of parametric and nonparametric models for traffic flow forecasting." Transportation Research Part C: Emerging Technologies 10(4): 303-321.
- [15] Li, L., Y. Li, et al. (2013). "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence." Transportation Research Part C: Emerging Technologies 34: 108-120.

Proceedings of the 21st International Conference on
Industrial Engineering and Engineering Management
2014

Qi, E.; Shen, J.; Dou, R. (Eds.)

2015, XXIII, 675 p. 405 illus., 100 illus. in color.,

Hardcover

ISBN: 978-94-6239-101-7

A product of Atlantis Press