Short-term Traffic Flow Prediction based on Data-Driven K-nearest neighbour Nonparametric Regression

Ken Chen¹, Shasha Zhao, and Dengyin Zhang²*
Nanjing University of Posts and Telecommunications University, Xinmofan Road, Nanjing, China
¹05761934@qq.com, ²zhangdy@njupt.edu.cn
*Corresponding author: zhangdy@njupt.edu.cn

Abstract. In order to solve the problem that the KNN algorithm is not accurate enough in the short-term traffic forecasting process and the prediction efficiency is not high caused by the search for past observations, a short-term traffic flow prediction model based on data-driven K-nearest neighbor non-parametric regression is proposed. The model is developed based on two-step search algorithm. Firstly, in the non-predictive period of time, the candidate input data similar to the current state is searched from the historical data for the system. Then, the optimal decision input data for prediction is identified from the candidate input data at the predictive point. Finally, the best decision input data is used to generate the prediction through the prediction algorithm. The simulation results show that the algorithm can effectively reduce the time for searching historical data on the premise of ensuring the accuracy of system prediction, thus reducing the execution time in the process of system prediction and improving the prediction efficiency of the system.

1. Introduction
Intelligent traffic control system has developed rapidly in recent years, and traffic flow forecasting has been widely studied as an important branch of traffic flow forecasting [1]. Traffic flow forecasting is to analyze data systematically based on collected multi-source traffic flow data and other influencing factors in traffic changes with randomness and uncertainty, and then find out the rules and construct corresponding predicting models and methods to predict future traffic flows [2,3]. Short-term traffic flow forecasting technology can generally be divided into two categories [4]. One is the traditional mathematical statistical algorithm model. The other is the model based on artificial intelligence technology, which is represented by BP neural network model, KNN algorithm model in machine learning and so on. However, due to the highly nonlinear and non-stationary characteristics of short-term traffic flow, the prediction model based on linear system theory is not ideal. The prediction model based on nonlinear system theory accords with the nonlinear characteristics of short-term traffic flow, so it has strong adaptability [5].

With the development and maturity of big data and data mining technology, Because of its simple structure and high computational efficiency, KNN has attracted more and more attention and research [6]. At the same time, Smith and Williams et al. made a guess as early as 2002 that the prediction results of nearest neighbor method may be further improved in a larger sample space [7]. Jinmin Chen proved the validity of KNN regression prediction on this basis [8]. However, in order to ensure the accuracy of prediction, KNN algorithm must keep enough feature samples, which will lead to low prediction efficiency.
In view of the above problems, this paper proposes a short-term traffic flow prediction model based on KNN. It proposes a two-step search algorithm to find and identify the best decision input data set of KNN algorithm model prediction from historical data. The results show that the two-step search method can effectively reduce the time of searching historical data and improve the prediction efficiency of the system on the premise of keeping the prediction efficiency of the KNN algorithm model.

2. Construction of Algorithmic Model

2.1. Algorithmic framework

Nonparametric regression is a data-driven heuristic prediction mechanism, which predicts future values by searching data similar to current observations in historical databases. Generally, it can be divided into three parts: building historical data, calculating similarity and designing prediction algorithm. Then, based on the new input, similar samples are found in the historical data, the samples are used as the input of the prediction algorithm, and the final prediction value is generated. Therefore, the key factors of short-term traffic flow prediction based on data-driven K-neighborhood nonparametric regression are selection of feature vectors and construction of prediction algorithm.

When non-parametric regression is used to predict short-term traffic flow, it is necessary to construct a representative historical database with large capacity. Ideal historical database contains various trends and typical laws of traffic state change. Current collected data can find similar trends, but redundant data will consume the running time of the algorithm. Therefore, it is particularly important to screen historical databases and identify wrong data. At the same time, we must also consider the adequacy of data to reflect as many traffic conditions as possible. After the construction of the historical database, the real-time traffic data can be predicted. The relevant elements of non-parametric model include feature vectors, distance measurement criteria, K value selection and prediction algorithm. Samples matching current real-time observation data can be found from the historical database. Finally, the traffic prediction quantity at the next moment can be obtained by using the prediction algorithm. The algorithm flow can be represented as follows:

![Figure 1. Flow chart of KNN nonparametric regression algorithm](image_url)

2.2. Defining eigenvectors
To compare the current observation data with the historical data, we need to define a comparison criterion, which is the description of the eigenvector. In the short-term traffic flow forecasting system, road occupancy, driving speed and weather conditions will affect the traffic flow at the next moment. Even the traffic data of the nearest neighboring time and the adjacent section will have an impact on the traffic flow at the next moment. Therefore, the reasonableness of state vector selection is directly related to the prediction accuracy.

There is no uniform criterion for the selection of feature vectors. If too many factors are considered in the selection of feature vectors, it will not improve the accuracy of prediction, but will lead to a longer running time of the algorithm.

A database is established based on traffic flow data collected by traffic management department. For all traffic flow data of a measured section 24 hours a day in a year, it is assumed that $N$ continuous traffic flow data $x_c(t)$ represents the current state vector. The formula is as follows:

$$x_c(t) = [q(t-(N-1)T), \ldots, q(t-nT), \ldots, q(t-T)]$$ (1)

Where $n = 0, 1, 2, \ldots, N-1$. Let the measurement time of each flow data be $T$, $t$ represents the current moment, $q(t)$ represents the traffic flow in the period $[t-T, t]$.

Random selection of $M$ historical state vectors $x_j(\tau)$ equal to the current state vector $x_c(t)$ dimension from the database as feature vectors. The expression of $x_j(\tau)$ is as follows:

$$x_j(\tau) = [q(\tau-(N-1)T), \ldots, q(\tau-nT), \ldots, q(\tau-T), q(\tau)]$$ (2)

Where $j = 0, 1, 2, \ldots, M-1$, and $\tau$ are observed at a historical time point ($\tau < t$), and $q(\tau)$ represents the traffic flow in the period $[\tau-T, \tau]$.

### 2.3. Determining distance metric criteria

Distance measures the matching degree between real-time data and sample database. In this paper, Euclidean distance is used to measure similarity, that is to calculate the sum of square deviation of each component of the eigenvector and the corresponding points in the historical database. The formula is as follows:

$$d(q^e_{i} - q^\tau_{i}) = \left| q^e_{i} - q^\tau_{i} \right| = \left| \sum_{0}^{(N-1)T} q(t) - q(\tau) \right|^2$$ (3)

Where $d(q^e_{i} - q^\tau_{i})$ is the Euclidean distance between the current eigenvector and the historical eigenvector, $q(t)$ is the traffic flow of the current period, $q(\tau)$ is the traffic flow of the historical period.

### 2.4. Selection of K value

The nearest neighbor value $K$ represents the number of nearest neighbors selected from the historical database. To a great extent, the selection of $K$ value is related to the specific situation of historical database. At present, there is no definite criterion to guide the selection of $K$. The existing literatures also aim at their experimental data. According to the results of different $K$ values, we can find a better $K$. But we can confirm that too large or too small $K$ values will affect the prediction accuracy.

### 2.5. Result prediction

**Step 1:** The Euclidean distance $u_j$ between the current state vector $x_c(t)$ and the eigenvector $x_j(\tau)$ is calculated by formula (3). The $M$ values are calculated as $u_0, u_1, \ldots, u_j, \ldots, u_{M-2}$, $u_{M-1}$ respectively.

$$u_j = d(q^e_{i} - q^\tau_{i}) = \left| q^e_{i} - q^\tau_{i} \right| = \left| \sum_{0}^{(N-1)T} q(t) - q(\tau) \right|^2$$ (4)

The average values of these $M$ Euclidean distances are calculated and recorded as $\bar{U}$. The formulas are as follows:
\[ \bar{u} = \frac{\sum_{j=0}^{M-1} u_j}{M} \]  

(5)

The above M Euclidean distance \( u_0, u_1, \ldots, u_j, \ldots, u_{M-2}, u_{M-1} \) Compared with \( \bar{u}, Y(Y < M) \) samples less than the mean \( \bar{u} \) were taken out and used as candidate input data set \( k_p^* \text{-NN} \).

**Step 2:** The Euclidean distance between the current state vector \( x(t) \) and Y samples is calculated again in the candidate input data set by formula (3), the values of Y Euclidean distances are \( u_0, u_1, \ldots, u_j, \ldots, u_{Y-2}, u_{Y-1} \) respectively. \( Z \) samples whose Euclidean distance values are less than the average \( \bar{u} \) are taken as the best input set \( k_q^* \text{-NN} \) for decision making.

**Step 3:** Based on the selected \( k_q^* \text{-NN} \), the forecasting algorithm is used to predict the traffic flow \( q(t+T) \) in the future [t, (t+T)] as follows:

\[ q(t+T) = \frac{\sum_{i=1}^{Z} q(t+1)/u_i}{\sum_{i=1}^{Z} (1/u_i)} \]

(6)

### 3. Simulation analysis

**3.1. Simulation data**

The data of this experiment come from Nanjing Traffic Administration. In this experiment, the urban roads in Nanjing are divided into four categories. They are high speed road, main road, secondary road and branch road. According to the collected data, four sections are selected, one for each type of road section. The period is from November 1, 2017 to November 30, 2017, a total of 30 days of data, of which the first 24 days of data as the training data set of the KNN prediction framework, and the next 6 days of data as the test data set. The forecast period is from 8 a.m. to 6 pm, one hour later each time. The forecast interval is 10 minutes.

![Traffic flow statistics of four different sections in Nanjing](image)

**Figure 2.** Traffic flow statistics of four different sections in Nanjing

**3.2. Simulation verification**

According to the above KNN prediction framework, the absolute MAPE (Mean Average Percentage Error) between the predicted and observed values is calculated. The formula is as follows:

\[ E_t^{MAPE} = \frac{|v_t^{obs} - v_t^{P}|}{v_t^{obs}} \]

(7)

Where \( v_t^{obs} \) represents the observed average velocity at t time and \( v_t^{P} \) represents the predicted average velocity at t time point. The figure 1 shows the KNN prediction framework, the two-step search KNN (KNN-S2S) framework and the real MAPE values from 8:00. to 18:00 every hour, as shown in Figure 3.
3.3. Comparison of Prediction Results

For high speed road, main road, Secondary road and branch road. KNN based on two-step search algorithm (KNN-S2S) and RBF neural network are used to predict short-term traffic on November 6 respectively. The experimental results show that the prediction results of the three algorithms can better represent the real-time traffic flow of the current road sections under two different road sections. Moreover, the KNN prediction algorithm based on two-step search algorithm has significantly improved the prediction efficiency. As shown in Figure 4.

4. Conclusion

This paper analyses the application of KNN algorithm in urban short-term traffic forecasting. It finds that compared with the traditional model based on mathematical statistics. KNN algorithm has good performance in dealing with sudden change and non-linearity of urban traffic flow due to its non-parametric regression characteristics. However, the long execution time of KNN prediction system leads to the decrease of prediction efficiency. To solve this problem, the two-step search algorithm proposed in this paper finds and identifies the best decision input data set from historical data through two similarity measures. The experimental results show that this method effectively improves the prediction efficiency of the system on the premise of guaranteeing the accuracy of the original prediction.
Acknowledgements
This work is partially supported by National Natural Science Foundation of China (No.61571241 and 61872423), Industry Prospective Primary Research & Development Plan of Jiangsu Province (No.BE2017111), the Ministry of Education-China Mobile Research Foundation (No. MCM20170205), the Scientific Research Foundation of the Higher Education Institutions of Jiangsu Province (No. 17KJB510043), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX18_0889), China.

Reference:
[1] Na Zhao, Jiabin Yuan, Han Xu. Overview of Intelligent Transportation System [J]. Computer Science, 2014, 41 (11): 7-11+45.
[2] Bachman Lu, qinShu, Guanghua Ma. Short-term traffic flow prediction based on multi-source traffic data fusion [J/OL]. Journal of Chongqing Jiao tong University (Natural Science Edition): 1-7 [2018-11-21].
[3] Wenham Luo, Bantian Dong, Zheng Wang. Short-term traffic flow prediction based on CNN-SVR hybrid in-depth learning model [J]. Transportation system engineering and information, 2017, 17 (05): 68-74.
[4] MAIT, GHOSH B. WILSON S. Multivariate short-term traffic flow forecasting using Bayesians vector autoregressive moving average mode[c]//TRB. Transportation Research Board 91st Annual Meeting. Washington DC: TRB,2012:3728-3740.
[5] Fichu Ma, Zhejiang Rao. Research on short-term traffic flow forecasting methods for urban roads [J]. Highway, 2017,62 (06): 192-196.
[6] Kaiya Fu. Short-term traffic forecasting of urban roads based on KNN algorithm [A]. Chinese Association of Science and Technology, Ministry of Transportation, Chinese Academy of Engineering. Papers of the World Transportation Congress 2018 [C]. 2018:12.
[7] Smith, Brian L., Williams, Billy M. & Oswald, R. K. Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting [J]. Transportation Research Part C, 2002, 10(4): 303-321.
[8] Jiangmen Chen. Short-term traffic flow prediction based on KNN regression [J]. Microcomputer application, 2015, 31 (09): 25-29+4-5.
[9] Giorgi Laszlo, Michael Kohler, Adam Krzyzak & Harro Walk. A Distribution-Free Theory of Nonparametric Regression [M]. New York: Springer-Verlag, 2002:4-96.
[10] Xiaofan Dou. KNN algorithm overview [J]. Communication World, 2018 (10): 273-274.