Research of intelligent traffic flow prediction algorithm

Min Tu1, Baoshan Liu1, Fangqiang Zhong1
1Jiangxi Police College, Nanchang, Jiangxi, 330100, China
1Nanchang University, Nanchang, Jiangxi, 330100, China
1Nanchang University, Nanchang, Jiangxi, 330100, China

Corresponding author and e-mail: Min Tu, 114371789@qq.com

Abstract. With the rapid development of transportation industry, how to carry out intelligent traffic flow prediction has become an urgent issue. In this paper, compared with the single algorithm traffic flow prediction model, we propose a new method for modelling that combines the minimum multiplication support vector machine and clustering algorithm to predict the traffic flow, which provides the basis for scientific traffic management and decision-making process.

1. Introduction

In recent decades, with the rapid development of national economy and transportation construction, the number of motor vehicles has surged sharply. That phenomenon has also led to more and more traffic problems and got an amount of concern from society. Both traffic accidents and traffic jams have brought great harm to the city and even the country. How to reduce the pressure and harm caused by traffic problems has become a social concern.

Foreign researchers began research on the aspect of how to effectively apply the principles and models of various traffic flow predictions to foresee short-term traffic flow around 1960, and achieved practical and considerable results [1]. In summary, in the study of short-term traffic forecasting, two modeling methods are used mainly: mathematical model is included in the first one and the other without a mathematical model. The former mainly includes time series algorithm, Kalman filter algorithm, historical mean algorithm, short-term prediction based on chaos theory [2]. The important algorithms without mathematical model include nonparametric regression algorithm, neural network algorithm and fuzzy neural network [3-5].

Compared with the research of short-term traffic flow prediction abroad, researches within China is still in a low exploratory stage in this field, and there is still a big gap in terms of the actual application.

2. Traffic flow data processing

The traffic flow data used in this paper is collected by the East Lake District Traffic Management Bureau of Nanchang. These data are the real traffic flow data in a certain period of time, with the characteristics of time series, so they are typical time series data.

The signal monitoring period of the road in Nanchang is generally controlled within 120 to 300 seconds. The longest period of the special intersection will not exceed 600 seconds. The general driving speed of vehicle is within 10~60km/h, Generally, the driving time in a monitoring road section is within 1-5 minutes.
The data used in this paper is from November 2013 to December 2013 in Nanchang. It contains multiple sections and multiple intersections. The data is the traffic of each intersection. The details are shown in the Figure 1.

![Figure 1](image_url)

The original data of traffic flow.

From Figure 1, we can see that the original data contains many attributes, and has different meanings.

| Serial number | Intersection number | Types of license plate | Colors of license plate | vehicle type | Number of Lane | Timeline | Vehicle speed |
|---------------|---------------------|------------------------|------------------------|--------------|----------------|----------|--------------|
| 743396323     | 360102000003        | 2                      | 2                      | 2            | 01             | 2013-12-20 00:00:00 | 52           |
| 743396457     | 360102000003        | 2                      | 2                      | 2            | 01             | 2013-12-20 00:00:00 | 44           |
| 743397675     | 360102000003        | 2                      | 2                      | 2            | 01             | 2013-12-20 00:00:00 | 42           |
| 743397662     | 360102000003        | 2                      | 2                      | 2            | 01             | 2013-12-20 00:00:00 | 40           |
| 743398405     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 42           |
| 743398419     | 360102000003        | 5                      | 5                      | 5            | 01             | 2013-12-20 00:00:00 | 56           |
| 743398433     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 40           |
| 743398801     | 360102000003        | 2                      | 2                      | 2            | 01             | 2013-12-20 00:00:00 | 48           |
| 743398946     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 41           |
| 743398996     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 40           |
| 743392127     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 45           |
| 743393212     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 37           |
| 743393227     | 360102000003        | 2                      | 2                      | 2            | 01             | 2013-12-20 00:00:00 | 55           |
| 743393285     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 36           |
| 743394684     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 37           |
| 743398909     | 360102000003        | 1                      | 1                      | 1            | 02             | 2013-12-20 00:00:00 | 40           |

Table 1. Description of field named intersection number.

Serial number | Intersection number
---|---
27 North Road south toward the North | 360102033003
27 North Road North to South | 360102033004
Rich and great road (Tangshan Road) south to North | 360102001003
Rich and great road (Tangshan Road) north to South | 360102001004
Nanjing West Road east to the West | 360102007001
Nanjing West Road West to East | 360102007002
The road of folk morality West to East | 360102006002
Fuzhou Road east to the West | 360102005001
Fuzhou Road West to the east | 360102005002
Fold mountain Road East to West | 360102004001
Fold mountain Road West to East | 360102004002
Xiangshan North Road South to North | 360102003003
Xiangshan North Road North to South | 360102003004
East section of yangming east road west | 360102029001
East section of yangming west road east | 360102029002
Along the river north road _ from south to north | 360121000201
Along the river north road _ from north to south | 360121000202
The 81 bridge from south to North | 360102002001
The 81 bridge from North to south | 360102002002

From table 1, we can see that there are as many as 20 values in the intersection number field. Similarly, the type of vehicle, the number of lane and other fields also have several fixed values. Each value has its own representative meaning. Because the useful fields used in this article only include time, speed, intersection number, other properties are not described in detail here.

In the research of this paper, the traffic flow information of each section is separately counted, and the traffic sections are individually predicted. Take the 27 North Road South to the north as an example: we select a day as the research object, through the data cleaning method, to obtain the section for 5 minutes for the whole day traffic flow, that is, the whole day 24 hours into the 24*12 segment traffic data.

We know that traffic flow data usually has the following important characteristics:
First, real-time. The requirement of this feature for traffic prediction is that the speed of response need to be fast.

Second, accuracy. The purpose of traffic flow prediction is to manage traffic and make corresponding decisions, so the accuracy of prediction is required to be high.

Third, the correlation between space and time. It can be understood as sequence in time and adjacency in space.

Since the least squares support vector machine has high calculation accuracy and fast calculation speed [6], it is suitable. The relevant characteristics of traffic flow prediction can meet its needs, so the minimum multiplication support vector machine can be used to model and predict traffic flow.

3. Establishment and experiment of prediction model.

3.1. Traffic clustering based on time segment segmentation
Traffic flow data is a kind of data with time series’ characteristics. The use of time series as a feature to segment traffic flow data is an important aspect of research on time series by domestic and foreign researchers in recent years. There are many research methods for time series segmentation. One of the most common methods is called piecewise linear description method. This method uses the corresponding linear model to segment the data of time series, and describes each segment accordingly. At present, there are many literatures describing this method.

There are other methods of time series segmentation, such as the method based on probability density function, which can be described by using different probability models for different sequence segments through different models.

There is a way to classify traffic flow data in a total of four modes. The data of each model is modeled by ARIMA model, then the models established by each mode and their mutual transfer mapping are studied to realize the prediction of traffic. This process is transfer ARIMA [7]. Although more than one ARIMA model has been established in this prediction method, and the model is used to describe the traffic flow data, there is no such method based on fundamental generation mechanism of time series data and the characteristics of the traffic flow data itself to predict the dynamic ARIMA.

This paper predicts traffic flow and adopts a method of clustering traffic flows with time series. The specific ideas of this method are as follows:

First, select the data of a certain day of a certain street as an example, calculate the flow value of each time period of the traffic data of the day, and then use the Euclidean distance as a distance function to perform data clustering of traffic flow, initially take k=4;

Secondly, through the first step, we can get 4 classes D1, D2, D3 and D4, then K-means clustering is performed for each class; we regroup D1 into 2 categories, D2 clusters into 3 classes, D3 clusters into 3 classes, and D4 clusters into 2 classes. This clustering is not based on traffic, but clustered by time. By doing this, all the data can be grouped into 10 classes, which have the characteristics of time series.

Thirdly, through this quadratic clustering method, 10 classes can be obtained, and they have time series characteristics and are divided into 10 segments in time. There is a corresponding data size difference between each segment.

Fourth, the centroid calculation and quality calculation are performed for each class. The corresponding time is taken as the abscissa of each centroid, and the mean of the traffic flow is taken as the ordinate of the centroid. Quality is the number of classes, then according to the size of the centroid, that is, the time sequence, the centroids are sorted respectively: T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, and the centroids corresponding to each centroid are also marked as: M1, M2, M3, M4, M5, M6, M7, M8, M9, M10;

Fifth, T1', T2', ... T9' are calculated according to the following formula 1.

\[
T1' = T1 + \left(\frac{M1}{M2 - M1}\right) \times (T2 - T1)
\]
3.2. Specific experiments of time segment segmentation

Previously, we analyzed how to use the quadratic clustering algorithm to cluster and segment traffic flows with time series characteristics. We take the east to west section of Dieshan road as an example, get the traffic flow data of December 20, 2013 as the experimental sample, realize the segmentation of the traffic flow data of that day, then get the corresponding centroid and quality:

\[
\begin{align*}
\text{centroid} & \quad \text{quality} \\
T1 &= 2:15 & M1 &= 38 \\
T2 &= 4:32 & M2 &= 22 \\
T3 &= 8:21 & M3 &= 35 \\
T4 &= 9:13 & M4 &= 45 \\
T5 &= 11:22 & M5 &= 28 \\
T6 &= 14:02 & M6 &= 32 \\
T7 &= 17:38 & M7 &= 12 \\
T8 &= 19:14 & M8 &= 20 \\
T9 &= 21:19 & M9 &= 27 \\
T10 &= 23:09 & M10 &= 19 \\
\end{align*}
\]

Then, use the formula (1) to get each split line:

\[
\begin{align*}
T1' &= 7:30 \\
T2' &= 8:32 \\
T3' &= 9:21 \\
T4' &= 11:13 \\
T5' &= 12:22 \\
T6' &= 14:02 \\
T7' &= 16:38 \\
T8' &= 19:14 \\
T9' &= 22:19 \\
\end{align*}
\]

The flow sequence of one day is divided into 10 segments by 9 dividing lines. From 22:21 to 24:00, and 0:00 to 7:30 are actually connected, so the day is actually divided into 9 segments. There is a relatively large difference in flow concentration for each time period.

3.3. Combination of minimum multiplication support vector machine and clustering algorithm

In the above, we use the Quadratic Clustering to segment traffic, we can use the least-multiplying support vector machine algorithm to achieve traffic prediction for nine segments of traffic after segmentation. In the specific experiment, each group of predictions has five data variables that are historical data before the predicted time points, which are the traffic flow in the first 5 minutes, the first 10 minutes, the first 15 minutes, the first 20 minutes and the first 25 minutes of the point. These five data variables are used to predict the flow at this point.

After prediction, the result of prediction can be obtained, and the predicted error probability is obtained. If the second clustering is not performed first, the least squares support vector machine is used to predict directly, and the prediction error rate is about 25.33%, and the quadratic clustering algorithm is used to segment and then model for each time period. It is predicted that the error rate will be reduced to about 20%.

Therefore, through the combination of secondary clustering and minimum multiplication support vector machine, we can use the traffic flow with time series characteristics. According to the size of the traffic and the characteristics of the time, we can carry out secondary clustering and divide each time period with similar flow in a day. From the results, we can see that there is a certain regularity in traffic flow during the day, with early peaks, peaks at noon and peaks in the evening. By dividing the flow of the day into 10 segments, we may get a more detailed traffic distribution. The distribution of traffic flow plays an important role in the management of traffic and according to this, the intelligent transportation system can conduct a certain degree of decision analysis and formulate certain policies to manage traffic. The prediction of traffic flow by the least-multiplying support vector machine can provide relatively
accurate predictions and give the traffic system better information support to make automatic adjustment of traffic signals and other flow control methods.

![Figure 2](image_url)

**Figure 2.** Prediction error rate after combining the two algorithms.

Figure 2 shows the prediction error rates for each time period of traffic flow prediction after combining the two algorithms and the detail data can be found in Table 2.

**Table 2.** Prediction error rate after combining the two algorithms.

| Time          | 0:00-7:30 | 7:30-8:23 | 8:23-9:32 | 9:32-11:43 | 11:4-12:57 |
|---------------|-----------|-----------|-----------|------------|------------|
| Error rate    | 19.2      | 18.87     | 19.56     | 20.54      | 21.13      |

| Time          | 12:57-14:46 | 14:46-16:28 | 16:28-19:34 | 19:34-22:21 | 22:21-24:00 |
|---------------|-------------|-------------|-------------|-------------|-------------|
| Error rate    | 20.41       | 19.08       | 19.86       | 22.35       | 20.34       |

Through experiments, it can be seen that the error rate can be controlled at about 20% in each time period, which proves that this prediction scheme has certain stability. Moreover, compared with the prediction of least-multiplying support vector machine, the error rate is obviously reduced, which indicates the feasibility and effectiveness of this scheme. Such prediction can provide more accurate traffic flow prediction, provide more valuable information for intelligent traffic management, and better meet the needs of traffic flow prediction and traffic management.

4. Conclusions

Studying the different forms of traffic flow and its potential laws, and then establishing a scientific, efficient and practical traffic flow prediction model is the key research direction in the field of intelligent transportation. This paper describes several important characteristics of traffic flow data and the corresponding data mining techniques and is focus on two algorithms (K-means algorithm and least squares support vector machine) and how to use these two algorithms for modeling.

We can get the useful data for traffic flow prediction by preprocessing some historical data of Donghu District in Nanchang. Compared with the traffic flow prediction model of single algorithm, a combined forecasting model is proposed that uses quadratic clustering and the minimum multiplication support vector. Based on the processed data of traffic, we can get a short-term prediction of traffic to the street (approximately 5 minutes - 15 minutes).

In this paper, the combination of two algorithms, quadratic clustering algorithm and least squares support vector machine, achieves effective prediction of traffic flow. However, due to some characteristics of traffic flow data, the complexity of the transportation system and the incompleteness of the data, there are still many imperfections, including the following:

1. The traffic flow information is complex, and it contains various connections. For example, the traffic conditions of the same road segment will be different in different time periods; different road segments will also exhibit inconsistent traffic flow conditions at the same time.
The flow characteristics, vehicle speed characteristics and occupancy characteristics of traffic flow data are all mutually restricting and influencing each other. In this topic, the flow forecast only analyzes the data of traffic flow, which is insufficient. Future research direction should take these three aspects into consideration, and build a relationship model between them to get more detailed prediction and analysis, which will be more reasonable and accurate. In the research of this paper, only the data from some road sections in Donghu District of Nanchang is analysed and predicted. Research results need to be extended to other regions.

There are no studies on various special road sections, such as highways, trails, or national roads, and these aspects need to be addressed. At the same time, the studied model does not consider unexpected factor, such as traffic accidents, abnormal weather. These aspects need to be improved and perfected in the future.

(3) Data mining is becoming more and more important nowadays, and the research of data mining technology needs more time and energy invested, so as to perfect and improve the used prediction algorithm and further improve the prediction effect.

Acknowledgement
The research is supported by the Open Project of the Collaborative Innovation Center for Economic Crime Detection and Prevention and Control Technology of Jiangxi Province (No. JXJZXTCX-031), Also thanks to Jiangxi Province for its key R&D project: Research on Electronic Data Trajectory Analysis Technology under Cloud Computing (No. 20161BBE50044)

References
[1] Yu Yong. Application of combined forecasting method in urban traffic flow prediction [J]. Journal of Chongqing University of technology and industry (NATURAL SCIENCE EDITION), 2007,24 (2): 122-125.
[2] Jin, Feng, and Shiliang Sun. "Neural network multitask learning for traffic flow forecasting." Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE InternationalJoint Conference on. IEEE, 2008.
[3] Wang Xiaoyuan, Zhang Jinglei, Wu Fang. Research on traffic flow data cleaning rules [J]. Computer Engineering, 2011.
[4] Lu Yun. Research and application of data mining in clustering method [D]. Hefei: Master's thesis of Anhui University, 2007.
[5] Li Zhonghao. Research and application of support vector regression machine [D]. Dalian: Doctoral Dissertation of Dalian University of technology, 2006.
[6] Zhang Chaoyuan, Hu Guanghua, Xu Tianze. Establishment and comparison of three traffic flow prediction models [J]. Journal of Kunming University of science and technology: Science and Technology Edition, 2006, (4): 104-107.
[7] Zhang Tao, Hu Lifang, Liu Zhongxin, and Zhang Yuejie. Nonparametric regression for the short-term traffic flow forecasting. In Mechanic Automation and Control Engineering [M], 2010 International Conference on, pages 2850 –2853, june 2010.
[8] Xiaoyan Gong,Three Improvements on KNN-NPR for Traffic Flow Forecasting,IEEE 5 ITS Conference,Singapore,2002.
[9] Lu Haiting. Short term traffic flow prediction based on SARIMA and SVM model [D]. Nanjing: Master's thesis of Southeast University, 2010.