Short-term Traffic Flow Prediction Based on Deep Learning
Dong-mei ZHAI, Chao-hui SHI and Hong ZHAO*

School of SoftWare, Beijing Jiaotong University, 100044, China

*Corresponding author

Keywords: Short-term traffic flow prediction, Traffic flow extraction, Convolutional neural network, Support vector regression.

Abstract. The accuracy of short-term traffic flow prediction is affected by two factors: one is the accuracy of data collection, and another is model selection. During data collection, aiming at the high cost and low precision of the traditional use of fixed equipment to collect traffic flow data, this paper proposes an algorithm model combining convolutional neural network and support vector regression, and designs an input matrix considering the influence degree of the road segment. The example is proved that the proposed algorithm model is better than the ARIMA and SVR model, and it is an effective traffic flow prediction method.

Introduction

Through flow is the number of cars passing through a specific area of a certain road surface within a specified time. Short-term traffic flow forecasting is the forecast of the number of vehicles passing through a region in the next 5-30 minutes, usually in 5 minute intervals. It can provide data support for intelligent transportation systems, and is the premise of real-time traffic signal control, traffic assignment, route planning, self-navigation, detecting accident, etc. in intelligent transportation systems.

In the past 30 years, scholars at home and abroad have organized extensive research on short-term traffic flow forecasting and have obtained rich research results. The research results can be summarized into five categories: prediction models based on statistical analysis, nonlinear theoretical models, simulation prediction models, intelligent prediction models, and hybrid models. Among them, the most extensive and in-depth study is the ARIMA model based on statistical analysis. Levin and Tsao used this method to predict traffic flow on expressways. Weng Xiaoxiong et al. analyzed the data of an intersection in Guangzhou and studied the parameter selection of the ARIMA model. With the deepening of machine learning theory research, the intelligent prediction model has also developed. Csatro-Neto and so on use Online-SVR for predicting the traffic flow. Li Yuanyuan et al. combined phase space reconstruction and SVR to predict short-term traffic flow. In addition, deep learning is also used in forecasting the short-term traffic widely. Luo Xianglong et al. used the differential removal trend, and then used the Deep Belief Network (DBN) to complete feature learning and SVR for prediction. Arief Koesdwiady et al. used DBN algorithm to study the influence of weather factors about traffic flow, and proved the effectiveness of the algorithm with actual data. Muhammad Arif et al. used deep learning and nonparametric regression to forecast traffic flow. Wu Y. et al. used a single dimension convolution network to get the spatial features of traffic flow, and used two methods of long-term and short-term memory to excavate the short-term periodicity of traffic flow, and designed a depth of traffic flow prediction system based on feature level fusion.

In terms of data acquisition: most of the data in these studies were collected using traditional fixed equipment, including magnetic frequency detectors, wave frequency detectors, video bayonet detectors, etc. The fixed equipment realizes the data collection work, which requires the purchase and installation of equipment in the early stage, and the detection cost is high. Moreover, it is easy to appear that the equipment is aging, malfunctioning, etc., which will result in inaccurate data collection.
collection and large errors. Aiming at these problems, this paper have proposed a traffic data extraction method based on GPS data map matching and manual selection of detection points, and named it as MM-MDP method. The method uses GPS data as a data source to achieve high-quality extraction of short-term traffic flow data through map matching, manual mark detection points, and traffic extraction. The method has the advantages of high precision, strong portability, free choice of detection point location, and low data acquisition cost.

In terms of traffic prediction: most of the models mentioned above cannot excavate the spatial features of traffic flow well, and the research on spatial features is relatively simple, that is, the data sources of these studies are mostly the traffic data of highways, between the detection points. Most of the permutations are linear. In order to make good use of the spatial information between traffic flows and provide a predictive model suitable for urban complex roads, A short-term traffic flow prediction model based on deep learning was presented in this paper, named MI-CCNNS. The model is designed to calculate the influence degree of the road segment, and then the input time-space matrix is reconstructed. Then the convolutional neural network is used for feature extraction. The extracted feature vector is input into the support vector regression model for prediction and prediction is obtained result. Since the convolutional neural network has the characteristics of weight sharing, the model can get the spatio-temporal features effectively of the traffic flow data while reducing the complexity of the model, reducing over-fitting and reducing the amount of calculation. The characteristics of local perception of convolutional neural networks can be combined with the degree of influence of road segments, which not only can improve the accuracy of prediction, but also make the prediction results more interpretable and provide a basis for traffic management.

**MI-CCNNS Model Design**

The MI-CCNNS model has two main points: one is the structure of the model; the other is the input matrix corresponding to the model. This section mainly describes these two parts.

**Model Structure**

The model uses a convolutional neural network for feature extraction and then uses a support vector regression model to achieve prediction. As shown in Figure 1, the convolutional neural network portion consists of two convolutional layers and a pooled layer and a full connection layer. The output of this part is the extracted feature. Then connect the SVR to achieve prediction and output the results.

Traffic flow prediction is generally considered to be a time series prediction problem. If we do not consider the spatial factor and define $x_t$ as the traffic flow value of the current section of the $t$ period, then the short-term traffic flow prediction uses $O = \{x_t | t = 1, 2, 3, \cdots, T\}$ to predict the traffic flow $x_{T+1}$, ie $x_{T+1} = f(O)$. The popular ARIMA model falls into this category. If the time-space factor is considered comprehensively, then let $X = (X_1, X_2, \cdots, X_N)$, where $X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,T})$, $N$ represents the number of upstream and downstream related sections of the detection point to be predicted, and $T$ represents the prediction of the $T+1$th period using the previous $T$ periods. Let $Y$ be the predicted value of the $T+1$ period of the road segment to be predicted, then $Y = f(X)$.

![Figure 1. MI-CCNNS model structure diagram.](image-url)
This model takes into account the time-space factor and uses CNN and SVR models to combine predictions. Therefore, for the MI-CCNNS model: if the detection point to be predicted is P, the input of the model is X and the output is Y. X and Y are shown in Equations 5 and 6, respectively.

\[ X = (X_1, X_2, \cdots, X_N)^T \]  \hspace{1cm} (1)

\[ Y = (x_{p,T+1}) \]  \hspace{1cm} (2)

Where \( X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,T}) \), \( i=1,2,\cdots, N \). N indicates that the detection point P has N upstream and downstream detection points, and T indicates that the data of the previous T periods is used to predict the T+1 period. Both N and T count from 1 onwards.

After inputting X into the CNN model for processing, the output is obtained as:

\[ H = \varphi(X) \]  \hspace{1cm} (3)

Where \( \varphi \) represents the operation content of the CNN model.

The feature vector H is input to the SVR model for calculation, and the output is:

\[ Y = f(H) \]  \hspace{1cm} (4)

Where f represents the processing of the feature vector by the SVR model. Y is \( x_{p,T+1} \), that is, the traffic flow predicted value of the P detection point in the T+1 period.

The specific forecasting process is:

Step 1: The traffic flow data is normalized according to the RSI and then normalized.

Step 2: Import data into the CNN model to extract features.

Step 3: Enter features into the SVR model for prediction.

Step 4: Denormalized to get the predicted value.

**Space-time Input Matrix Based on Road Segment Impact Degree (RSI)**

The traffic flow at the detection point P is affected by time and space factors. So its input should take into account both time and space. Therefore, the input of the model can be constructed as a two-dimensional matrix, the horizontal axis direction of the matrix is arranged in time, and the vertical axis direction is arranged in the geographical position of the detection point.

Different models correspond to different input formats. If the input format is better, the prediction result of the algorithm model can be made more accurate. Therefore, because the convolutional neural network has the characteristics of local perception, this paper considers that the order of the detection points in the matrix will affect the prediction results, and proposes the concept of “degree of influence of the road segment” as the basis for the arrangement of detection points.

For a road segment, it may have multiple upstream or multiple downstream, and the influence of different upstream and downstream road segments is different. “Risk Impact Degree (RSI)” is a quantitative expression of this.
As shown in Figure 2, it is assumed that the future traffic flow of the detection point a is predicted, and the degree of influence of the detection point b on a needs to be calculated. The calculation manner is as follows: the flow rate of the detection point a in the time period t is $X_{a,t}$, using $X_{a,b,t}$ to represent the value from b (to b) in the flow of the detection point a in the time period t, then the degree of influence of b on a is called RSI, and the RSI can be showed as:

$$RSI = \sum_{t=96}^{120} \left( \frac{x_{a,b,t}}{x_{a,t}} \right) \frac{24}{t=96}$$

Where the t is the sequence number of the time cycle, and t=1 is the time range from 0:00 to 0:05 on a certain day, at intervals of 5 minutes, and t is continuously accumulated. Therefore, t=96 in Equation 4-3 represents the time range of 8:00-8:05, t=120 represents the time range of 9:55-10:00, and 24 represents 8:00 to 10:00, a total of 24 to 5 Minutes are interval data. The reason for choosing t is 96-120 is: 8:00 am to 10:00 am is the early peak time, the traffic volume is large, the data is more accurate, and the forecasting is more concerned about the peak of travel, so this time period is selected in the formula. The calculation of the influence coefficient between the detection points can make the calculation result more accurate.

After the RSI is calculated, the predicted detection point data is placed in the first row, and each detection point is arranged from near to far according to the size of the RSI. The larger the RSI is, the closer the first row is, and the input matrix is obtained.

This paper believes that when predicting the traffic flow value of the detection point P in the T+1 period, the traffic flow value of P on the previous day and the previous day's T+1 time can be used as a reference. Then, if the prediction is performed at intervals of 5 minutes, one day can be divided into 288 time intervals. Therefore, in the sample construction, $x_{p,T+1}$, $x_{p,T+1-288}$ and $x_{p,T+1-576}$ are added to the last two columns of the input matrix of the sample and get the final input matrix.

Then, N detection points are selected, and the detection point P is predicted using the flow rate of the first T time periods of the current time period. The input part of a sample should be:

$$X_{N\times(T+2)} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1T} & x_{1T+1-288} & x_{1T+1-576} \\ x_{21} & x_{22} & \cdots & x_{2T} & x_{2T+1-288} & x_{2T+1-576} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NT} & x_{NT+1-288} & x_{NT+1-576} \end{bmatrix}$$

The corresponding prediction result is $x_{p,T+1}$. 

184
MI-CCNNS Experimental Design and Results Analysis

Data

The MM-MDP model is a process of converting GPS data into traffic data, providing a data foundation for the MI-CCNNS model and one of the influencing factors of the accuracy about predicting. The data source of the MM-MDP model is the GPS data provided by the Drip Trip. The model is a process from GPS trajectory data to flow data. The data is processed through four steps: data preprocessing, map matching, manual mark detection, and traffic extraction:

(1) The data preprocessing process needs to process the missing data and the abnormal data. This article chooses to delete it.

(2) Map matching is implemented using an improved location-based approach. This method parses each GPS point using the Google Geocoding API to get additional data Location Type, Viewport, and bounds. The Location Type can have four different values. According to the value of the Location Type, different matching algorithms are used respectively, and finally the matching result is obtained. When the Location Type is ROOFTOP or GEOMETRIC_CENTER, the resolved location coordinates are the corrected coordinates. When the Location Type is APPROXIMATE, the corrected coordinates are calculated using the vertical projection method. The vertical projection method uses the two sets of southwest and northeast values that are resolved, that is, the two points are formed into line segments, and the coordinate points are projected onto the line segment, which is the corrected coordinate point [11].

(3) The detection point does not actually exist on the real road, but exists in the electronic map, it has no entity, just a kind of data. The method of setting the manual detection point is: after completing the map matching, according to the map information, selecting the road segment to be monitored and setting the detection point. The latitude and longitude data of the detection point is obtained by means of a tool such as a high-tech open platform, and finally the coordinate information of the detection point (rectangle) is obtained.

(4) After the manual marking of the detection point is completed, the traffic flow data needs to be extracted. This process is mainly based on the latitude and longitude and the divided time period, the traffic flow data is extracted from the GPS trajectory data.

The flow data obtained by using the MI-MDP model is a data source for sample preparation. After correlation coefficient analysis, it is found that the trend of traffic changes on weekends and non-weekends is different, so this paper predicts non-weekend data. The traffic flow time interval in this paper is set to 5 minutes, and a total of 7092 samples are obtained, using 80% as the training set and 20% test set. There are 5,676 training sets and 1,418 test sets.

Evaluation Indicators

The model’s evaluation indicators use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The detailed calculation of these two indicators is shown in Equations 9 and 10:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|^2}
\]  
\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|
\]  

In the formula, \(x_i\) represents the actual traffic flow value for the i-th period, \(y_i\) represents the traffic flow value predicted by the same time period model, and \(N\) represents the total number of predicted time periods.
Experimental Parameter Setting and Result Analysis

The experimental parameters are designed according to the experience and the model and input format of the paper. The specific parameters of the CNN model are: the first layer is causal convolution, the size of convolution kernel is $5 \times 5$, the step size is 1, and there are 16 convolution kernels. The second layer adopts an average pooling, and the first layer of the pooling layer has a size of $2 \times 2$ and a step size of 2. The second layer of convolution is a causal convolution with a convolution kernel size of $5 \times 5$, a step size of 1, and a number of 16. Then connect the second layer of the pooling layer, using an average pooling, the size is $2 \times 2$, the step size is 2. The model iteration number is 10000; the learning rate is 0.001; the batch-size is 100; to prevent over-fitting, the dropout layer is introduced with a parameter of 0.2; using Adam optimization. The SVR layer uses a radial basis (RBF) kernel function with a penalty factor of $C = 0.1$ and $\varepsilon = 0.1$. The sample input part matrix size is $14 \times 14$.

The experimental environment is 2.60GHz, 4 cores, Intel(R) Core(TM) i7-6500U, 8G memory. Use Jupyter as the development environment, develop the language as Python, and use the TensorFlow framework.

In order to verify the validity of the model, this paper selects the ARIMA model and the SVR model for comparison. The experimental value is the average of 50 calculation results. The prediction results of the three models are shown in Figures 3, 4 and 5. The performance indicators are shown in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI-CCNNS</td>
<td>8.1</td>
<td>9.8</td>
</tr>
<tr>
<td>ARIMA</td>
<td>13.73</td>
<td>18.82</td>
</tr>
<tr>
<td>SVR</td>
<td>10.2</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Table 1. ARIMA, SVR, MI-CCNNS prediction results comparison table.

![Figure 3. MI-CCNNS comparison of predicted and actual values.](image)
Comparing Figures 3, 4 and 5, the MI-CCNNS model proposed in this paper has better prediction effect, higher degree of fitting, and MAE and RMSE are superior to ARIMA model and SVR model, which proves the MI-CCNNS proposed in this paper. The prediction method is an effective short-term traffic flow prediction method.

Conclusion

A new method of traffic flow acquisition and a short-term traffic flow prediction method using deep learning were presented in this paper. The traffic flow acquisition method provides an accurate and reliable data source for the prediction method. The new traffic data acquisition method converts GPS data into traffic data by means of map matching and manual selection of detection points. The predictive model uses flow data as input, uses convolutional neural networks for feature learning, mines the essential characteristics of the data, and then uses support vector regression for traffic prediction. Through the example verification, the model proposed in this paper has a good prediction effect.

References


