A Short-term Traffic Flow Prediction Approach of Neural Network Based on Cluster Analysis

Zhengwu Yuan*, Weiwei Zhang & Ming Yang
College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

ABSTRACT: Given the principle of different dates having different traffic flow patterns, the research clusters the historical data based on the results of clustering, predicts the class, which flow pattern the date is in, and then outputs all data of this pattern to forecast the date’s traffic flow. On the choice of clustering method, we uses k-means algorithm. Since the k-means algorithm is not unity in the choice of the optimal clustering number, we combine KNN algorithm to determine the optimal cluster number k, then use it as the classification method, and choose BP neural network to predict short-term traffic flow. The experiment proved that the approach is better than that of without clustering analysis in prediction accuracy and the method of k-means algorithm combining with KNN algorithm to determine the optimal cluster number k is effective.

Keywords: cluster analysis; K-means; KNN; BP neural network; short-term traffic flow forecasting

1 INTRODUCTION

Traffic jam is a big problem of the society nowadays and it’s also an important issue in ITS (Intelligent Transportation System). Transportation planning or control is the key to establish perfect intelligent transportation system and it can solve the problem of traffic congestion, realize traffic management and guidance. To achieve transportation planning, it needs to forecast the traffic flow accurately, real-time and quickly is of vital significance to guide people’s travel.

In general, predicting traffic flow has two ways: macroscopic prediction and microscopic prediction. The former is to get a lot of traffic data and to predict traffic flow by the year, month or day and it can divide holidays and non-holidays; the latter establishes a unified database with all the traffic flow data for short-term traffic flow prediction and predicts the traffic flow of a certain moment in a road. Lots of scholars research this approach, but few cluster and analyze the data first. The paper uses clustering algorithm to analyze the characteristics of traffic flow based on a data of a year in a row, classifies them respectively with classification algorithm and according to the category, and uses the neural network algorithm for short-term traffic flow prediction.

2 LITERATURE REVIEW

Short-term traffic flow refers to that the time interval of traffic flow prediction ranges from 5 to 15 minutes. It can be roughly divided into two categories: traditional statistical models and Computational Intelligence (CI) approaches [1]. The former is a method based on the traditional theory, such as historical average model, time series analysis model, Kalman filtering model, Autoregressive Integrated Moving Average (ARIMA), etc. The latter is a prediction method based on computational intelligence model, including nonparametric regression, neural network, support vector machine (SVM), etc.

Historical Average Model [2] is based on the cyclical characteristics of traffic flow, using average of the history data to predict. Average model assumes that the traffic flow changes periodically over time. It is a kind of static prediction method, relying on the traffic volume of historical moments in the road to be forecast. It’s an approach with simple computing. In the calculation functions, volume to be predicted presents a linear relationship with the history volume, but the
actual traffic volume and forecast flow have no linear relationship. The method is not suitable for nonlinear transport system, for it has certain difficulty in solving the emergency in traffic.

Time series model [3] is a method by processing dynamic random data through the parameterized model. It is suitable for dynamic prediction with a simple calculation and can be updated by real time data. However the method is not suitable for short time traffic flow prediction.

Regression analysis prediction method [4] presupposes causality between the research object, and through the mutual influence between the object to predict traffic flow of the future time. So it is often used to analyze multiply road section traffic. So it needs a variety of traffic data in many sections. But this method invalids when the data is insufficient.

Kalman filter theory [5] is an advanced method widely adopted in modern control theory. It is flexible and can deal with both stationary and non-stationary data, applying to linear, non-real-time and online traffic flow prediction, which has high prediction accuracy for steady traffic flow. But the kalman filtering method needs cumulative values in computing every time and it’s more complicated and sometimes it has a delay in predicting the outputting value. Since it is a linear model, it can’t predict the traffic flow data of short time interval. With the augmentation of the short-term traffic flow nonlinearity, uncertainly, the performance of this model becomes poor.

In recent years, traffic forecasting methods have been gradually shifting from traditional statistical models to CI approaches [6]. For the traffic system is a complex and nonlinear time-varying system with the combination of human being, vehicles and road and it has much uncertainty causing from the season and the climate. Neural network [7] can recognize the characteristics of complex nonlinear system. Therefore it is more suitable for short-term traffic flow forecasting. Neural network has advantages of self-learning and adaptive. So it can update the network according to the real-time traffic information and can guarantee real-time prediction. It not only can take advantages of the historical data of sections to be researched, but also consider the influence of related sections and all kinds of factors in traffic system, such as weather conditions, road construction, accident, road conditions, etc.

3 METHODOLOGY

People’s travel demand has vast differences in different dates, leading to volume in the distribution of different date having different features. Such as on weekdays, people’s travel has a fixed time for work and study. That’s the reason why there will be an early peak and a late peak in a day. However, during the holidays, people’s entertainment travel demands increase, while the travel demands of study and work reduce greatly. So the holiday’s traffic is different from that of weekdays [8].

As shown in Figure 1 and 2, traffic flow data of detector 1045 of a high-speed beltway in the twin cities, Minnesota, USA, on December 25, 2015, which is Christmas day, traffic peak appeared in 12:00 am-20:00 pm. On December 12, 2015, which is a common workday, traffic peak appeared at 07:00 pm-09:00 pm and 14:00 am-15:30 am and the value of peak is different in non-workdays and workdays. In addition, from the traffic value in the two graphs we can also find the detector 1045 is in section of where the local people go to work or study place rather than the entertainment area.

If traffic flow data of holidays is used to predict that of working day, there must be a big error. It needs to separate different traffic patterns, and flow pattern library is needed to be set up respectively. So how to distinguish the flow pattern will have a great effect on the accuracy of traffic flow prediction. And experience alone can’t find the differences between different traffic patterns accurately and efficiently, so it needs to adopt the method of cluster analysis to distinguish the similarity of traffic flow.
high-speed beltway of the twin cities, Minnesota, from December 11, 2014 to December 11, 2015. According to the result of clustering, we build different flow pattern bases for every class, then classify real-time data, select the class of pattern base, and finally obtain the data in the repository and use the neural network algorithm to forecast. The whole technical route is as follows (See Figure 3):

![Diagram](image)

**Figure 3. The technical route of this research.**

### 3.1 Clustering algorithm

Clustering is to cluster the data into different classes, making the similarities of similar cluster larger and that of heterogeneous clusters smaller. The current clustering algorithms can be roughly divided into the hierarchical clustering algorithm, partition clustering algorithm, clustering algorithm based on density, clustering algorithm based on grid and clustering algorithm based on the model, etc. One of the most commonly used is K-means algorithm, which belongs to the partition clustering algorithm, for its simple implementation, high efficiency on large data.

Steps of K-means algorithm are as follows:

1. Select from k data randomly from the n data as the initial clustering center;
2. Calculate the distance of the rest of the (n-k) data and the k data, and make the smallest distance data gathered to one class;
3. Calculating the average of new clustering and make it be a new clustering center;
4. Repeat steps 2 and 3, until the clustering center is not changed.

K-means algorithm needs to specify the number of clusters k first, and the value of k directly affects the result of clustering. Under normal circumstances, the value of k needs to be specified by users according to their own experience and understanding of the data set. When the value of K is not specified so appropriate, results of clustering also can’t be guaranteed. When the value of k is too small, it will lead to the differences in the data of class and do not reflect the difference between real data; while the value of K is too big, it can’t find the common features of different categories. In traffic flow prediction, all kinds of date in different traffic time series is our research object. This paper chooses the Euclidean distance between time series as a measure of similarity degree. The calculation method is as follows,

\[
d_j = \sqrt{\sum_{i=1}^{N} (X_{1i}^2 - X_{2i}^2)}
\]

In the formula (1), i is the number of moments to in one day, j is the number of days, \(X_{1i}\) and \(X_{2i}\) refer to the volume of class 1 and 2 at the i moment, \(d_j\) refers to the Euclidean distance of class 1 and class 2 at the i moment on the date of j.

In addition, the clustering effect of K-means can’t be determined by the macroscopic observation and experience. It is not only time consuming, but also making the results not be optimal. So it needs to have a standard to measure clustering effect stand or fall, namely the validity of clustering. Single standard often has some limitations, and it needs to measure a variety of standards to achieve the optimal clustering effect. So this research combines with external and internal standard to determine the optimal clustering number, and makes it a way to measure the pros and cons of clustering results. The distance D, which is a distance between each point and all kinds of clustering center, is selected as an internal standard and classification accuracy namely A as an external standard. D goes smaller, indicating that the distance between different classes smaller; A goes larger showing that clustering is more effective. But due to the reality of traffic flow characteristics and the need of prediction, the value of K can’t be too large. However, the classification here is not known before clustering, but after classifying. The effect of clustering is measured by the classification accuracy in turn.

### 3.2 Neural network

Neural network \(^9\), which was created in the 1940s, is a biological process to simulate the brain intelligent technology. Neural network has a variety of forms, and Back Propagation BP (Back Propagation) neural network\(^10\) is most widely used.

BP neural network is a one-way transmission of multi-layer forward neural network, even with 3 layers or more than 3 layers of network structure. Three-layer network structure refers to the input layer, hidden layer and output layer. In the theory, it has been proved that the three-layer network model has good effect on nonlinear modeling.

### 3.3 Measuring error between traffic profiles

The method to measure error is called the absolute percentage error. Absolute percentage error refers to the percentage of the absolute values of the differences between the predicted values and real values,
reflecting the forecast the degree of deviating from the observed value or real value. The formula is as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

(2)

In formula (2), the actual value is $y_i$, the predictive value is $\hat{y}_i$ and the total of samples for forecasting is n.

4 METHODOLOGY ADOPTED IN THIS RESEARCH

4.1 Data set

Select data from 4500 detectors distributed in high-speed beltway in the twin cities. All detectors were sampling data every 30 seconds, every day per week and continuous sampling. This study selects the No.100 detector, from December 11, 2014 to December 11, 2015. Data is integrated by every 5 minutes, in a total of 105408 moments.

4.2 Clustering results analysis

K-means clustering algorithm is used for clustering analysis with the MATLAB R2013b. As mentioned before, the value of K can’t be too big and to make sure that the classification accuracy is above 80%, take K = 2 ~ 15. When running the algorithm combined with K-means and KNN, the result of the optimal cluster number K is as follows (See Figure 4):

As is shown in Figure 4, when k=8, the result is best. The entire data set divided into 8 classes is the best cluster. The average of every class at every moment is drawn in the Figure 5.

Figure 5. Average volume at each moment

4.3 The prediction results and analysis

Different flow patterns of data are obtained by clustering analysis. KNN algorithm is used to predict the date of its categories and BP neural network to predict its value. Then we get the comparison of forecasted values and measure it with actual values. December 7, 2015 (Monday, not holiday) is selected as the research object, separately forecasted by clustering analysis of BP neural network and without cluster analysis at 6:00am-21:00pm. The comparison of predicted values and actual values figures (Figure 6, Figure 7), and the error graphs (Figure 8, Figure 9) are shown as follows:
As is shown in those figures, after clustering analysis, prediction of traffic flow on December 7, 2015, and the average absolute error is 0.11543 and without clustering analysis is 0.22161 respectively. After cluster analysis, the error is decreased obviously. Therefore, neural network prediction after clustering analysis improves the accuracy of the prediction.

From the two graphs in Figure 6 and Figure 8, it can easily see that the accuracy in Figure 6 is higher and from Figure 7, errors at some moments are even below 5. Besides, in the Figure 6, the track of prediction is more similar to the track of the actual value than that of Figure 8 and the moments, whose errors are below 5, are more. Therefore, it indicates that the algorithm of combination of K-means and KNN to determine K is feasible and neural network after cluster analysis can improve the accuracy of traffic flow prediction.

5 CONCLUSIONS

This paper presents an approach of neural network based on clustering analysis in predicting short-term traffic flow. That is, before the short-term traffic flow prediction using clustering to distinguish traffic patterns. According to the result of clustering, create different databases, choose the appropriate flow database, and predict traffic volume with BP neural network algorithm. What’s more, after clustering and classifying, it can easily know which class the date to be researched is in, which flow pattern it is and when the peak will come approximately, which has a great significance for traffic control and guidance. For example, according to the research, the Christmas Day of 2015 is in class 3, and we can forecast the moments of traffic jam is about 10:30-16:30 and then we can avoid going outside in this period. Results show that the cluster analysis to classify traffic patterns makes average absolute error decrease and prediction accuracy high. Based on this study, the further research can be on the similarity of the track of the time series and by this way, the prediction accuracy of traffic flow can be improved.

6 ACKNOWLEDGEMENT

The author would like to appreciate the funding support from the National Natural Science Foundation of China, Chongqing and the Regional Transportation Management Center (RTMC).

REFERENCES

[1] Jing Liu. Wei Guan. 2004. Traffic flow forecasting method review. Highway Traffic Science and Technology, 21(3): 82-85.
[2] Yun Lan. 2002. Short-term Traffic Flow Prediction Research. Xi’an: Northwestern Polytechnic University.
[3] Zhongjun Liang. 2010. Short-term Traffic Flow Prediction Method Based on Correlation Analysis of Time and Space Research. Chongqing: Chongqing University of Posts and Telecommunications.
[4] Smith B.L., Dermetsky M.J. 1997. Traffic flow forecasting: comparison of modeling approaches. *Journal of Transportation Engineering, 123*(4): 261-266.

[5] Habtemichael, F.G., Cetin, M. 2015. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transport. Res. Part C*.

[6] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, Yunpeng Wang. 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C, 54*: 187–197.

[7] Dougherty M S, Cobbell M R. 1997. Short-term inter-urban traffic forecasting using neural networks. *International Journal of Forecasting, (13)*: 21-31.

[8] Yang Liu, Shoufeng Ma. 2013. Based on the clustering analysis of nonparametric regression short-term traffic flow forecasting method, *Traffic Information and Security, 2* (31): 27-31.

[9] Bingzheng Xu, Bailing zhang, Gang Wei. 1994. *Neural Network Theory and Application*. Guangzhou: South China University of Technology Press.

[10] Hongjie Wang, Liangming Lin, Dazha Xu et al. 2001. Dynamic temporal prediction algorithm based on improved BP network traffic flow study. *Journal of Traffic and Computers, 19*(99): 13-16.