Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

Community Traffic State Prediction Based on CPM and LSTM

Yuange Ma *
North China Electric Power University, Baoding, 071000, China.
*Corresponding Author.

Abstract:

Accurate prediction of road traffic status can effectively alleviate the increasingly serious urban traffic congestion and promote the development of intelligent transportation system (ITS). Different from other studies that predicted only one road area, this paper predicted a group of road areas with traffic similarity (a community), which made it easier and more accurate to find the key features. Based on CPM algorithm, LSTM model and FCM algorithm, this paper predicts the overall speed and flow of the community and divides it into some traffic states, effectively captures the long and short time characteristics of time series, solves the limitations of traditional road prediction research and improves the prediction accuracy. In order to verify the accuracy and effectiveness of the model, we selected taxi data in New York City for model training and testing. The results show that the model can predict the speed and flow of community 1and divide it into some traffic states well, with good prediction accuracy, and has great application value in the road traffic state prediction.

Keywords*: Traffic state prediction; Clique percolation method; Long Short Term Memory Network; Fuzzy C-means clustering; Complex network theory.*

I. INTRODUCTION

In the face of the continuous increase in the number of motor vehicles, especially the number of private cars, while bringing a lot of convenience to the public, it also increases the burden of urban traffic roads. Due to the limitations of urban space resources, it is difficult to meet the expanding public demand by building a large number of new roads or widening old roads. As a result, the urban road traffic congestion problem is sharply highlighted, and its harms are as follows: For the public, traffic congestion brings great inconvenience to people's travel, making people have to pay far more time cost than expected to complete the travel, reducing people's travel efficiency. [1-2]For the state and government units, road congestion has brought huge time cost, and the inefficient urban operation efficiency has brought great negative impact on economic development and urban development [3]In addition, as traffic congestion increases travel time, the resulting energy consumption and exhaust pollution will also increase.At the same time, the low speed of vehicles will also cause incomplete combustion of energy, making environmental pollution more serious.[4]In general, traffic congestion not only seriously affects People's Daily travel activities, reduces the efficiency of social activities and causes huge economic losses, but also wastes a lot of resources and affects the development of cities.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

Intelligent Transportation System (ITS) using advanced information technology has become the preferred method to solve the traffic congestion in many cities at home and abroad. In the face of complex urban traffic environment, one of the core keys to realize intelligent urban traffic is the state perception judgment and prediction of traffic flow.By constructing the short-time traffic state prediction mechanism of urban road, the problem of urban traffic congestion can be effectively alleviated, the utilization rate of urban road resources can be improved, the traffic pressure of urban congested sections can be alleviated, and the air pollution caused by it can be alleviated.2. Accurate and effective traffic flow prediction will also provide basic data support for urban traffic signal coordination and control and trip time prediction, which is an important technical support for active traffic information service and scientific urban traffic management.[5]

Therefore, under the situation of sharply prominent traffic congestion problem and booming development of smart transportation city, accurate prediction of overall traffic state of urban community by using traffic big data will help alleviate traffic congestion problem and promote the development and construction of smart city. Based on this background, this paper proposes a method to predict the overall speed and flow of the community and divide it into some traffic states based on CPM algorithm, LSTM model and FCM algorithm. Taking New York City as an example, this paper first divides communities: The adjacency map within each administrative region was first obtained, and then the CPM faction filtering topology was used to divide each region within each administrative region into several communities. Each community contained some areas with traffic similarity, and then the data of average speed was sorted and assigned to the community. Second, predict the community speed and flow: After processing the data, it is added into the LSTM neural network to train and predict the community speed. And the community traffic data is obtained by processing the data, which is added to the LSTM neural network for training, and the community traffic is predicted. Third, divide the traffic state of the community: Taking community speed and community flow as parameters, FCM algorithm is used to classify and finally predict the traffic state of the community.

The contributions of this paper can be summarized as follows:

(1) This paper proposes a community overall traffic state prediction model combining CPM algorithm, LSTM neural network and FCM algorithm.

(2) Compared with the prediction of single road area, the prediction of the overall traffic state of the community can be easier and more accurate to find the key features, which plays a great role in alleviating urban traffic congestion, avoiding congested communities and choosing unblocked routes through map app navigation, and allocating the orders of ride-hailing drivers to ride-hailing cars located in unblocked communities, etc.

The structure of this article is as follows. Section 2 gives a literature review of traffic state prediction methods. Section 3 describes the data set and preprocesses the data. Section 4 introduces CPM faction filtering algorithm, LSTM neural network model and FCM algorithm. Section 5 discusses the results of the

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

case study, validates the prediction methods proposed in the study, and compares the prediction effectiveness of different methods. Section 6 draws the conclusion and looks into the future work.

II. LITERATURE REVIEW

With the rapid development of urban traffic, the imbalance between road supply and demand has become increasingly prominent, the traffic operation situation is inefficient, and traffic congestion has become an important factor hindering urban development.[6]Intelligent Transportation Systems (ITS) and machine learning algorithms can improve the comprehensiveness of traffic operation situation assessment system. More real-time and accurate prediction of traffic state plays an important role in optimizing the intelligent control system of urban traffic.

In terms of traffic state assessment, Krause traffic state recognition based on fuzzy logic, analysis of traffic flow and the speed of the car, the traffic state is divided into six grades, Yang Qingfang et al. discriminated and evaluated the traffic movement situation through fuzzy C-means clustering analysis, Huang Yanguo et al. pointed out that the classification and evaluation of traffic operation situation based on machine learning can improve the real-time evaluation of urban traffic operation situation. [7-9]

In terms of traffic flow prediction, scholars have put forward many traffic flow prediction models and methods. The commonly used prediction models can be roughly divided into traditional prediction model and neural network prediction model.[10]In the former, Ahmaed et al. first applied the autoregressiveintegral moving mean method (AIRMA) model to the prediction of expressway traffic flow. Mascha et al. applied ARIMA model to the study of short-term traffic flow prediction, and also considered the correlation of traffic flow time series.[11-12]Brian et al. used seasonal ARIMA model to predict short-term traffic flow on expressways and compared the advantages and disadvantages of prediction effects at different time intervals.[13]Gary et al. used the time series model to predict the changing trend of expressway traffic flow, and the experiment proved that the model has good robustness, that is, it can be applied in a small amount of discontinuous traffic flow data.[14]GUO et al. used the stochastic adaptive Kalman filter model to predict traffic flow and achieved good results.[15]For the traditional model, due to the inherent nonlinear and non-stationary characteristics of traffic flow, the prediction accuracy is often reduced and the anti-interference ability of the model is poor. Artificial neural network is very good at processing big data, plus its strong learning ability and adaptive ability, is widely used in traffic prediction. Yao Jialin et al. studied using K Nearest Neighbor (KNN) algorithm to predict the speed of each road section in the road network under the condition of missing data.[16]Zhang Li et al. proposed an improved BP neural network algorithm based on rough set and genetic algorithm, which improved the training accuracy and generalization ability of the model.[17]Dong Chunjiao et al. redivided the road network at the spatial level, reconstructed the traffic flow time series, used the new sequence as the input vector, and realized the simultaneous prediction of multiple sections of the road network with Elman neural network.[18]Li Qiaoru et al. constructed a more accurate traffic flow prediction model by integrating support vector machine (SVM) with spatio-temporal data. Zhao Yaping et al. proposed a traffic flow prediction model based on least squares support vector machine, and verified the advantages of fast

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

learning speed, good tracking performance and strong generalization ability through examples. After summarizing basic data to construct feature vectors of traffic flow and determining four prediction states, Tan Juan et al. used deep learning self-coding network to conduct tagless training and finally, Softmax regression model was built on the top layer to carry out polymorphic prediction of traffic congestion conditions.[19-21]Luo Xianglong et al. used difference technology to eliminate trend directions in traffic flow data and combined deep belief network and support vector regression for prediction. Nicholas et al. proposed a new deep learning structure for traffic flow prediction, which includes L1 regularized linear model and a series of tanh network layers.[22-23]For the traditional recursive neural network, it can not train the time series with long time interval, and there is gradient explosion or disappearance.In order to solve these shortcomings, Hochreiter et al. proposed LSTM NEURAL Network (LSTM NN). MA et al. used LSTM to predict traffic speed and found that it could capture the time features in time series, which was suitable for traffic prediction and had good performance in traffic prediction.[24-25]Deep learning is in a period of rapid development and has been widely used in the field of short-term traffic flow prediction.

It is not difficult to find that the research mentioned above has certain effectiveness in both traffic state prediction and road network prediction with combined temporal and spatial features. However, these studies do not take into account the temporal and spatial similarity characteristics of overall traffic flows in communities containing the selected predicted road areas. The prediction of only one road area will be subject to a lot of interference. In addition to the periodic rule of time characteristics, the influence of surrounding road sections should be considered in space, so it is difficult to accurately predict. However, if some areas with similar traffic are aggregated into a community, the overall prediction of speed, traffic flow, etc., will have a great advantage. First, in terms of time dimension, these regions all have similar time features which make it easier and more accurate to find time features; Secondly, in terms of spatial dimension, we do not need to explore the influence of surrounding space from only one road area, but can find the key characteristics of the overall traffic flow of this community more easily and accurately from the traffic flow of many areas with similar traffic characteristics. In general, instead of predicting one road area, predicting a group of road areas (a community) makes it easier and more accurate to find key features. At the same time, compared with the prediction of single road area, the prediction of the overall traffic state of the community plays a great role in alleviating urban traffic congestion, avoiding congested communities by navigation through map app and choosing unblocked routes, and allocating orders of ride-hailing drivers to ride-hailing cars located in unblocked communities. Therefore, based on Based on CPM algorithm, LSTM model and FCM algorithm, this paper proposes Community traffic state prediction model that predicts the overall speed and flow of the community and divides it into some traffic states.

TABLE 1: Model comparison

Traffic state assessment	Existing research		[7], [8], [9], this paper	
Traffic flow forecast	Traditional prediction model		[11] - [15]	The prediction accuracy is reduced and the anti-interference ability of the model is poor
	Artificial neural network	Traditional recursive neural network	[16] - [23]	Time series with long intervals cannot be trained, and there are gradient explosions or vanishes
		LSTM, GRU, etc	[24] - [25], this paper	Long-term time signatures can be captured and the problem of gradient explosion or disappearance solved

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

III. DATA PREPARING

3.1 Data Preprocessing and Topology building

The data of this experiment are from the order data of New York taxis from January 2016 to June 2020, including green taxis and yellow taxis. PULocationID (PUID), DOLocationID(DOID), LPEP_pickup_dateTime (PUT), lpep_dropoff_datetime(DOT), Trip_distance passenger_count (PC), (D), etc. The following is the partial order data of Green taxis in New York City in June 2020, as shown in Table 2:

TABLE 2: Dataset display

Index	PUID	DOID	PUT	DOT	PC	D
1	264	264	2019/12/18 15:52:30	2019/12/18 15:54:39	5	0
2	66	65	2020/1/1 0:45:58	2020/1/1 0:56:39	2	1.28
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3027	75	75	2020/1/1 4:06:19	2020/1/1 4:06:28	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

When we checked the data, we found some errors, so we first removed the following incorrect data:

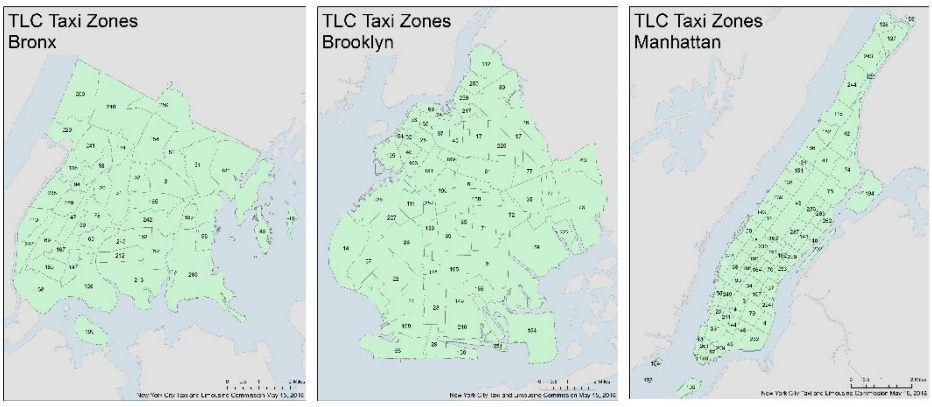
(1) Data of boarding and alighting time difference ≤ 0;

(2) Data with mileage ≤ 0;

(3) Data with Passenger capacity ≤ 0;

(4) Data that does not fall within the specified time range.

Since there is no direct speed and flow data in the data used, and in order to select areas with traffic similarity and integrate them into a community for prediction, this paper will establish the topological structure of the five administrative regions respectively according to the map of the five administrative regions of New York.



Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

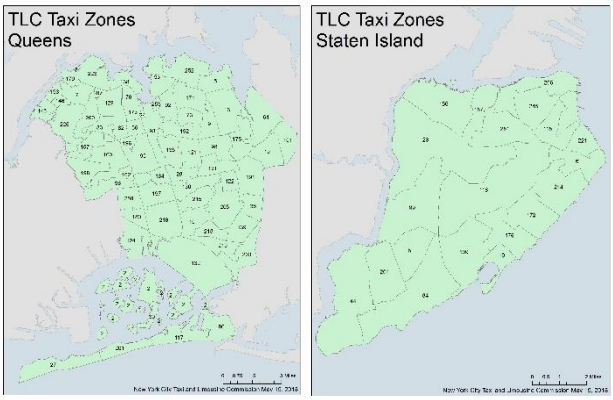


Figure 1: Five Boroughs 'abstract maps of New York

Meanwhile, in order to facilitate data preprocessing, we agreed as follows:

(1) The traffic flow in adjacent areas is directly related, while the traffic flow in non-adjacent areas is indirectly related. Considering that traffic flow prediction needs to extract spatial features, this paper decides to take the directly related region as the traffic similarity region.

(2) Since most vehicles flow only in one administrative region, this paper assumes that the traffic flow of each administrative region is independent and does not affect each other, that is, taxis are operated only in the administrative region.

(3) Considering the driving characteristics of taxis, this paper approximates the region passed by the shortest path in the corresponding administrative region topology as the region passed by the taxi starting and ending.

(4) For A single sample, this paper approximates the average speed of taxis from area A to area B as the current speed of all regions passed by the shortest path from the starting point to the end point.

(5) It is assumed that the average speed of taxi stays basically unchanged at C within 1 minute before and after the current time.

The following figure shows an example of establishing the topology of Statenisland:

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

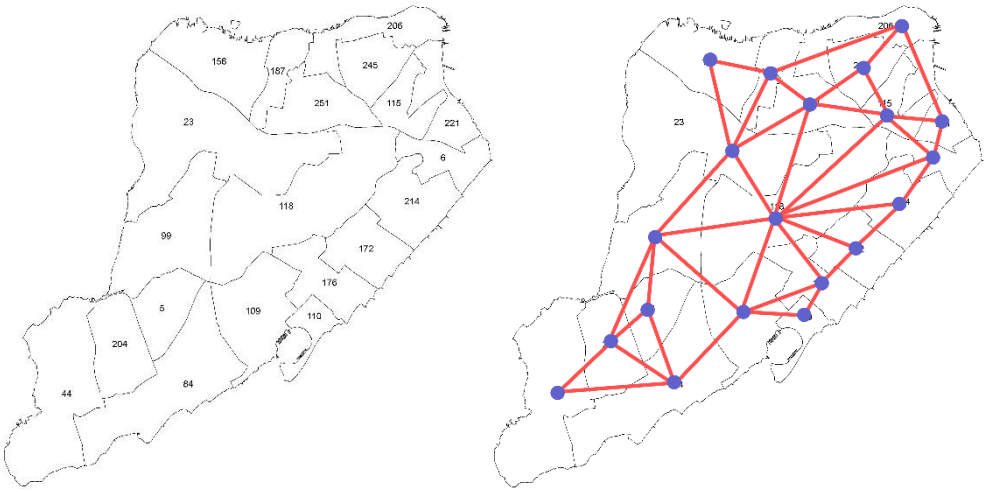


Figure 2: An example of constructing a topology of certain borough

3.2 floyd multi-source shortest path algorithm to find the shortest path

After the topological structure of different administrative regions is established, the set of shortest path nodes between two regions in the same administrative region is required to prepare for the following speed and flow prediction. For this problem, Floyd multi-source shortest path algorithm is used in this paper to solve the shortest path.

IV. METHODOLOGY

Considering that most of the current studies only predict the traffic flow of a single road region, and do not take into account the actual state of the overall regional traffic flow prediction requirements, this paper proposes a method based on CPM algorithm, LSTM model and FCM algorithm to predict the overall speed and flow of the community and classify the traffic state. CPM algorithm, LSTM and FCM algorithm are introduced as follows:

4.1 K-Clique Percolation Method of complex networks

The Clique Percolation Method (CPM) proposed by G.Palla et al is a classic overlapping community discovery algorithm, which is widely used in random networks, weighted networks, directed networks, binary graphs and other complex networks for community discovery based on Clique penetration theory[26-28]

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

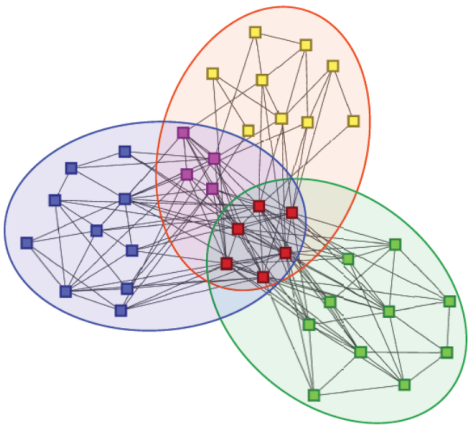


Figure 3 Complex networks structure

4.1.1 Definition of K-faction module

Palla et al. believe that a module can be regarded as a set of interconnected "small total coupled networks" in a sense. These "fully coupled networks" are called "factions", and k-factions represent the number of nodes in the fully coupled network as K. If two K-factions have k-1 common nodes, the two k-factions are said to be adjacent. If a k-faction can reach another K-faction through several neighboring K-factions, the two K-factions are said to be interconnected. In this sense, k-factions in the network can be regarded as a set composed of all connected K-factions. In a network, some nodes may be nodes in multiple K-factions that are contiguous (there is no K-1 common node). Therefore, these nodes will be "overlapping" parts of different K-faction modules. Therefore, CPM algorithm is usually used to find overlapping module structures in the network.

4.1.2 Finding factions on the Network

In the CPM algorithm, the algorithm of iterative regression from large to small is used to find the cliques in the network. First, from the degree of each node in the network, we can judge the size s of the maximum possible coupling network in the network. Start from a node in the network, find all factions containing the node of size S, delete the node and its connected edge; Then, select another node and repeat the above steps until there are no nodes in the network . At this point, all factions of size S in the network are found; Then, decrease S step by step (each time S decreases by 1), and repeat the above method; Finally, all factions of different sizes in the network are found.

As can be seen from the above steps, the most critical problem in the algorithm is how to start from a node V to find all the factions of size S containing it. For this problem, CP algorithm adopts iterative regression algorithm.

Firstly, for node V, two sets A and B are defined, where A is the set of all points connected in pairs including node V, and B is the set of nodes connected with all nodes in A. In order to avoid repeated selection of A node, nodes in set A and B are arranged according to the sequence number of nodes in the

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

algorithm.

On the basis of defining sets A and B, the algorithm is as follows:

(1) The initial set A={v}, B={v neighbor};

(2) Move A node from B to set A, adjust set B at the same time, delete nodes in B that are no longer connected to all nodes in A

(3) If the set B is empty before the size of A reaches S, or A and B are subsets of A larger faction, the calculation will be stopped and the previous step of recursion will be returned; Otherwise, when A reaches S, it gets A new faction, records that faction, and then returns to the previous step of the recursion to continue looking for the new faction.

From this, we have all the factions of size S starting from v.

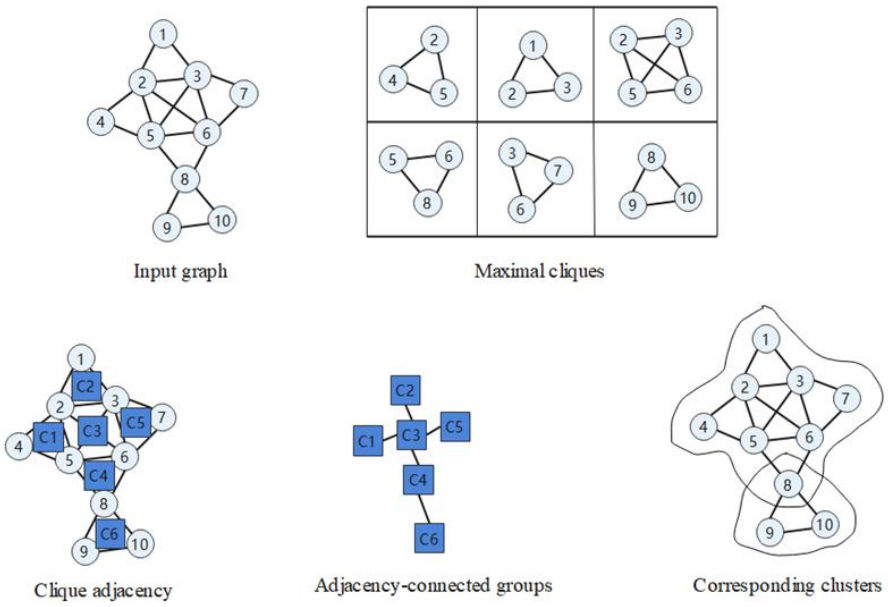


Figure 4 A step-by-step view of the original clique percolation method

In 3, there are many complete subgraphs in the administrative topological network structure constructed according to the convention, namely, the edge-dense network, which is suitable for the k-faction filtering algorithm (CPM) in the community division algorithm in the complex network. Therefore, this paper uses this algorithm to carry out regional aggregation in the administrative traffic area network. Taking the topological network division of Brooklyn as an example, the region aggregation results are shown in Table3, and the visualization results are shown in Figure 5. Nodes of the same color belong to the same community, nodes of different colors belong to different communities. It can be seen that CPM algorithm has good community division ability for the network structure with overlapping parts,

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

that is, a single node may belong to multiple communities, and the community division result obtained by CPM algorithm is more consistent with the actual situation of the network with overlapping parts in real scenes.

TABLE 3: The community partition results of Brooklyn obtained by using CPM algorithm.

Community	District set	Q
1	256, 97, 34, 36, 37, 80, 17, 49, 112, 181, 189, 217, 61, 25, 255	
2	256, 65, 34, 97, 33, 66, 17, 49, 217, 25	
3	257, 67, 227, 228, 133, 11, 14, 111, 21, 22, 181, 89, 26	
4	67, 11, 14	
5	225, 35, 37, 76, 77, 17, 177, 61, 63	
6	257, 178, 21, 181, 89, 26, 190	
7	257, 165, 71, 72, 91, 190, 178, 21, 149, 85, 89, 123, 188, 62	0.95942
8	108, 155, 210, 21, 150, 149, 55, 154, 123, 29	
9	65, 25, 33	
10	33, 228, 40, 106, 52, 181, 25	
11	33, 195, 40, 52, 54, 25	
12	35, 39, 72, 155, 76, 91, 222	
13	35, 39, 72, 49, 61, 181, 62, 91, 188, 189, 190	

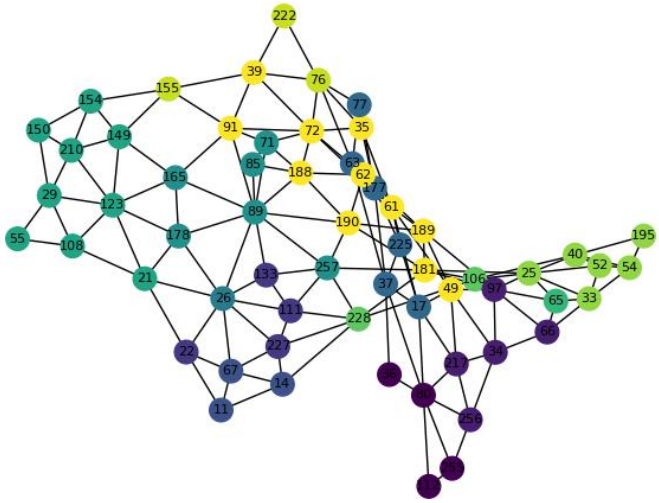


Figure 5: CPM result of Brooklyn

According to CPM algorithm, the visualization results of community division of topological network structure in the other four administrative districts of New York are shown in Figure 6.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

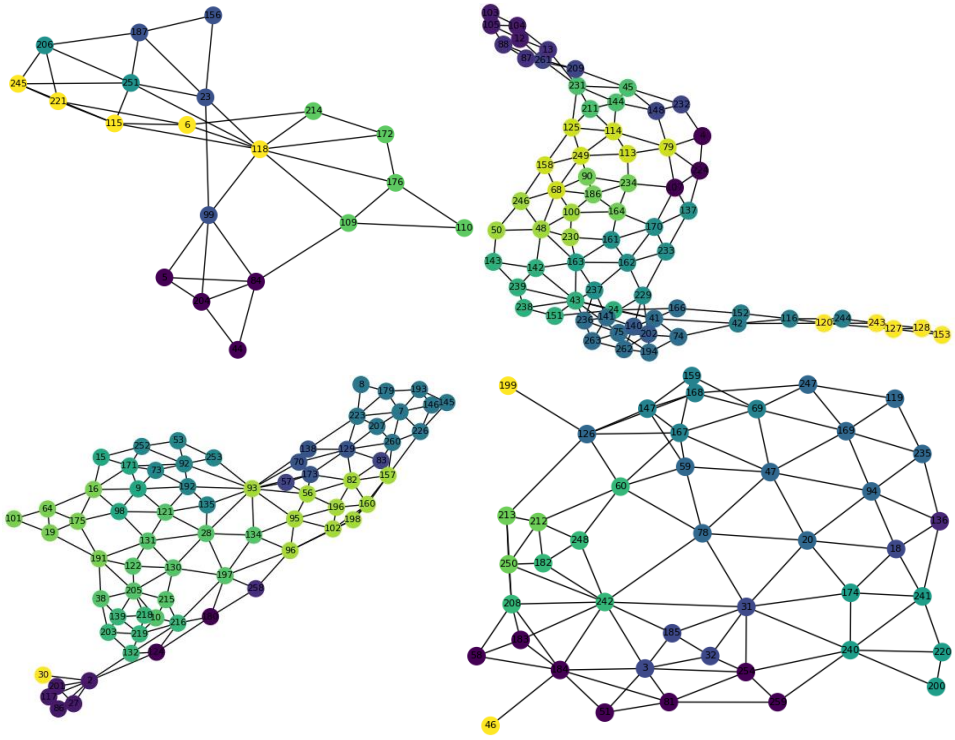


Figure 6: CPM results of Statenisland, Manhattan, Queens and Bronx

4.2 Long short-term Memory neural network

4.2.1 LSTM

LSTM model, also known as long short-term memory network, is a variant of recurrent neural network (RNN).[29]Recurrent neural network (RNN) is good at processing time-continuous data series, which are highly correlated before and after, while BP neural network and convolutional neural network do not associate information before and after time. Recurrent neural network (RNN) is a time-cyclic network, which allows information to persist. The network structure of RNN is shown in Figure 7.[30-32]

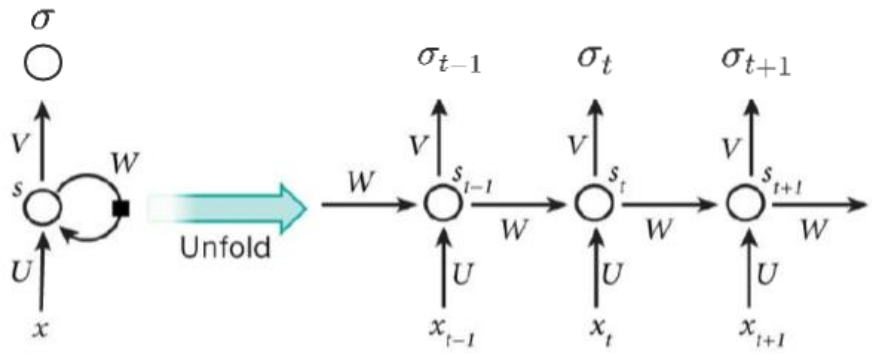


Figure 7 : Rnn structure

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

In deep learning, recurrent neural network (RNN) can better deal with time series problems. The output of RNN is determined by the current characteristic input and the state of the previous moment.Suppose a series of time-continuous behavioral actions can be expressed as $x=\{x_1, x_2, ..., x_t, ..., x_T\}$, x_t represents the t th behavioral action, and T represents the number of behavioral actions. Then the output of RNN hidden layer h_t is shown in formula (1).

$$h_t = \sigma_h(w_{xh}x_t + w_{hh}h_{t-1} + b_h) \tag{1}$$

In formula (1), σ_h is the activation function, w_{xh} is the weight matrix between the input layer and the hidden layer, w_{hh} is the weight matrix between the hidden layer and the hidden layer, h_{t-1} represents the state at the last moment of RNN, b_h represents the bias.

Then the output results of input layer of RNN are shown in Formula (2).

$$y_t = \sigma_y(w_{ho}h_t + b_o) \tag{2}$$

In formula (2), σ_y is the activation function, w_{ho} is the weight matrix of the hidden layer and the output layer, h_t b_o is the hidden layer output, is the output layer bias.

As the RNN model lengthens with time series, the gradient will disappear when the model gets deeper. To solve this problem, LSTM model is introduced, which is a variant of RNN and can store memory for a long time.The basic unit of LSTM consists of one cell and three gates, namely Forget Gate, Input gate and Output gate, as shown in Figure 8:[33]

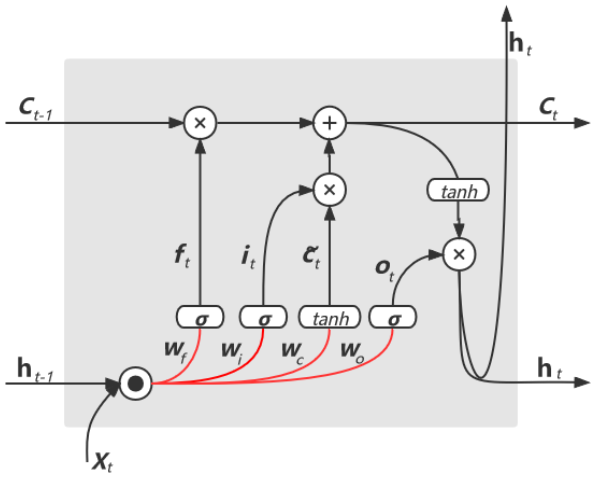


Figure 8 The structure of the LSTM

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

(1) Forget Gate: In this control unit, the inputh_{t-1} is the output of the last LSTM and the input x_t. at time T It can be expressed by the following formula:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \#(3)$$

(2) Input Gate: This part is different from the previous step, which is mainly to retain effective information in the calculation, as shown in Figure 8. It can be seen from the figure that the Input Gate is mainly to determine which data will be stored and transmitted backward, and whether to update the cell state.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \#(4)$$

In addition, as shown in Figure 8, there is also a tanh function for the input to update the old unit state, so that it is superimposed with the former. Given the outputh_{t-1} and input x_t transmitted at the previous moment, it can be obtained:

$$\tilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C) \#(5)$$

Finally, it can be determined that the unit state after updating at time T can be expressed as:

$$C_t = f_t \cdot C_{t+1} + i_t \cdot \tilde{C}_t \#(6)$$

As for this, the calculation of LSTM at time T is completed.The last step requires adding a base unit to pass the current calculation to the LSTM at the next time t+1.

(3) Output Gate: The input of the Output Gate is the final unit stateC_t . Tanh is used in this paper to determine which data will be transmitted further down.On the other hand, using sigmoId to determine the current cell state is printed:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (7)$$

By integrating the above two expressions, the input :h_t at the next moment can be obtained as follows

$$h_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \cdot \tanh(C_t) \quad (8)$$

As a method of deep learning, LSTM is suitable for solving the time series input and output problems.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

Since road traffic flow has periodic rules in time, LSTM is very suitable for road traffic flow prediction.In this paper, LSTM is used to predict the overall flow of the community.

4.3 Fuzzy C-means clustering

As a representative method of Fuzzy clustering analysis, Fuzzy c-means clustering (FCM) are: the basic principle of calculating the membership degree is used to measure the degree of sample data belonging to one another, using the iteration method to calculate all the classes of clustering center, to obtain the similarity index as the objective function is minimum.

4.3.1 Mathematical description

Given the traffic flow data sample set vector X, which has n elements, $X=\{x_1,x_2,x_3,...,x_n,\}$,the idea of Fuzzy C-means clustering algorithm is to divide sample set X into W clusters, $C=\{C_1,C_2,C_3,...,C_w,\}$, where, C_i represents the i th cluster cluster. The objective function is

$$\min(J_m(U,X,C))=\min(\sum_{i=1}^{n}\sum_{j=1}^{w}u_{ij}^{m}\parallel x_i-c_j\parallel^2) \tag{9}$$

Constraint conditions are:

$$\begin{cases}\sum_{j=1}^{w}u_{ij}=1,1\leqslant i\leqslant n\\ u_{ij}\in[0,1],1\leqslant i\leqslant n,1\leqslant j\leqslant w\\ 0<\sum_{i=1}^{n}u_{ij}<n,1\leqslant j\leqslant w\end{cases} \tag{10}$$

Where, c_j is the clustering center of the j th clustering cluster C_j (j∈ [1, W]), which is essentially the mean value of the corresponding clustering cluster data;

u_{ij}— represents the membership degree of the sample point x_i to the j th cluster C_j (j ∈ [1, W]), which should meet the constraint conditions shown in the formula above;

U=[u_{ij}]— represents membership degree matrix, where 1≤ i ≤n, 1≤j≤w;

m (m≥1) -- represents the fuzzy weighting index, which can be used to control the influence of membership degree . When m increases, the membership degree matrix will also increase, and the value

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

range of m is [1.5, 2.5].

And c_j and u_{ij} can be calculated by introducing Lagrange coefficient.

$$c_j = \sum_{i=1}^{n} (u_{ij})^m \cdot x_i / \sum_{i=1}^{n} (u_{ij})^m, \; j = 1, 2, 3 \ldots w \tag{11}$$

$$u_{ij} = \frac{1}{(\sum_{t=1}^{k} d_{ij} / d_{it})^{2/(m-1)}} \tag{12}$$

In Formula (12), d_{ij} represents the Euclidean distance between the sample point x_i and the cluster center c_j of the j th cluster C_j.

4.3.2 Algorithm principle and process analysis

FCM algorithm mainly realizes the clustering and classification of traffic state through two modules of fuzzy clustering and classification decision, and fuzzy clustering module can get four types of traffic.

The data set of state provides a relative standard for state discrimination of traffic flow operation data.The classification decision module can use the optimal clustering center to distinguish the traffic state attributes of a group of traffic flow data measured by the principle of minimum Euclidean distance.Therefore, fuzzy C means can realize the identification of traffic state.

Detailed FCM clustering algorithm process steps include the following parts:

Step 1: Set the initial parameters.Including clustering number w (w=4 in this paper), fuzzy weighting index m, convergence conditions,iteration number $T_{\max}$, iteration termination threshold ε;

Step 2: Select the initial clustering centers $c^0 = \{c_1^0, c_2^0, c_3^0, \ldots, c_w^0\}$ randomly and calculate the initial membership matrix U_0;

Step 3: Calculate the clustering center $c = \{c_1, c_2, c_3, \ldots, c_w\}$ and update the membership matrix U;

Step4: calculate the objective function $\| c^{t+1} - c^t \| \leqslant \varepsilon$. If the maximum number of iterations is reached, terminate the calculation; otherwise, return to step3 to continue the iteration..

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

The specific algorithm flow is shown in Figure 9:

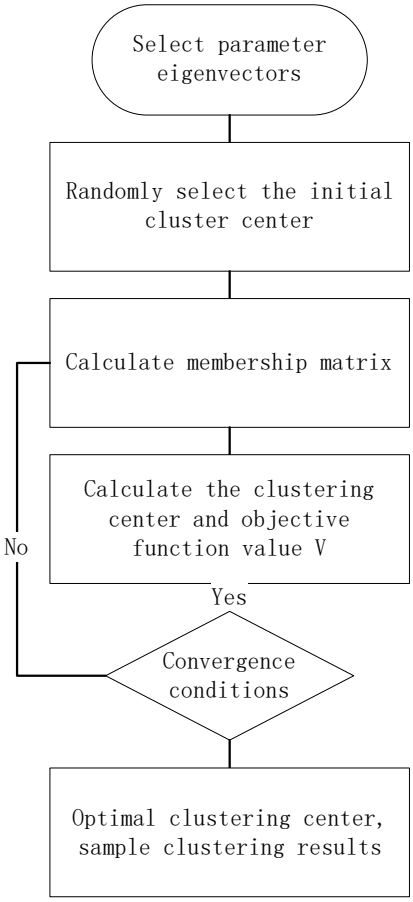


Figure 9 Algorithm flow

V. EXPERIMENT

The deep learning compiling environment used in this experiment is Pycharm, supported by Anaconda3 Data Science Platform, while deep learning Framework we adopted is Tensorflow2.3.0, Along with the corresponding CUDA version is 10.1.

TABLE 4: Deep learing environment information.

Compiling Environment	Python version	Deep Learning Framework	CUDA version
Pycharm	3.8	Tensorflow2.3.0	10.1

The experiment process is shown in Figure10;

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

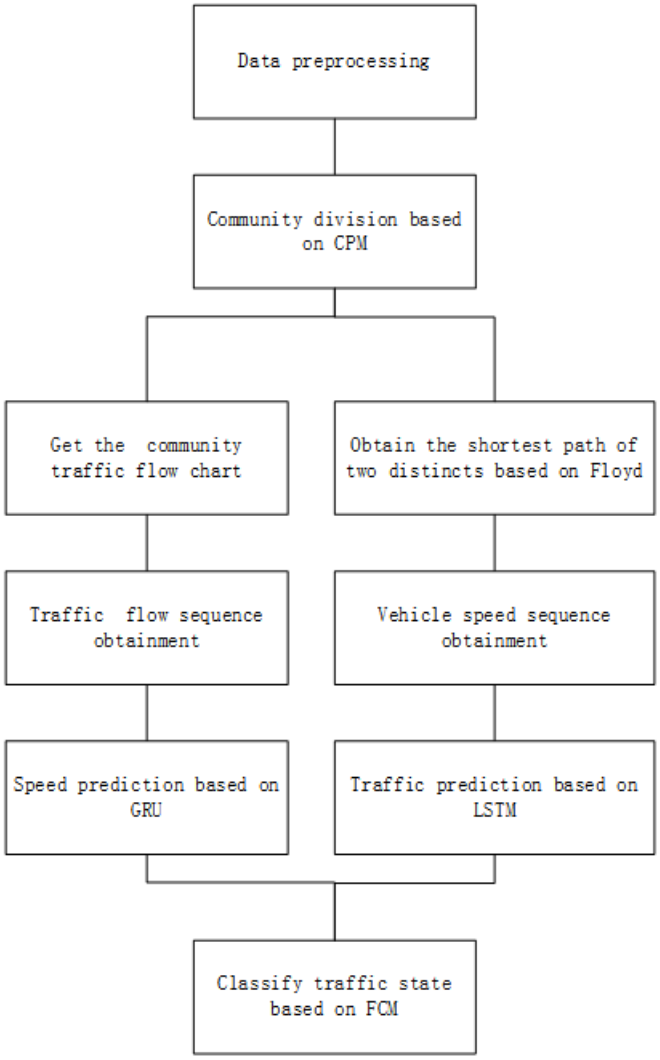


Figure 10. Experiment Process

5.1 Speed prediction based on LSTM

5.1.1 Data processing

In 3.2, k-faction filtering is carried out on the topology map constructed by the five administrative districts of New York, and some areas with similar traffic characteristics in each administrative district are bobbed into communities. Floyd algorithm is used to obtain the shortest circuit matrix P in each administrative district. According to the agreement made in 3.1, In this paper, the average speed of each sample is assigned to the region set containing the shortest circuit between the loading and unloading locations according to the topological network, and the above vehicle time is taken as the occurrence time of the average speed.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

Next, this paper will select communities for speed prediction.Due to the differences in the number of regions and data volume in the five administrative regions, this paper comprehensively considered the results of community division, sample size and number of regions, and decided to select Brooklyn administrative region for further processing and prediction.It is known that all the areas in Brooklyn are divided into 13 communities according to CPM algorithm. This paper takes community 1 as an example to predict community speed.

Next, this paper made the data set of vehicle speed time sequence.The first step is to discretize the sequence in the time domain.When setting the sampling interval, this paper took into account the application scenarios of this model, such as the prevention and control of traffic congestion, which requires real-time prediction. Therefore, in order to reduce the time cost, this paper decided to adopt the sampling interval of 15 minutes, taking into account both the speed and accuracy of prediction.

5.1.2 Speed prediction based on LSTM

The construction of LSTM in this paper is shown in algorithm 1. The number of LSTM units in the first layer is 32, step size =15, feature number =3, and the activation function is hyperbolic tangent function. A Dropout layer of 0.5 has been added to prevent overfitting. The number of LSTM units in the second layer is 16. The mean square error was used to measure the difference between the predicted value and the actual value during training, and Adam was used as the optimizer.

Algorithm1

```
model = keras.Sequential([
tf.keras.layers.LSTM(32, activation='tanh', return_sequences=True, input_shape=(15, 3))),
tf.keras.layers.Dropout(0.5),
tf.keras.layers.LSTM(16, activation ='tanh'),
tf.keras.layers.Dense(1),
])
model.compile(loss='mean_squared_error', optimizer ='adam')
```

In this paper, the speed of community 1 is divided into training set test set by 8:2, and the training results are shown in the figure. It can be seen from the figure that the model in this paper has achieved high prediction accuracy, which proves the feasibility of community speed prediction

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

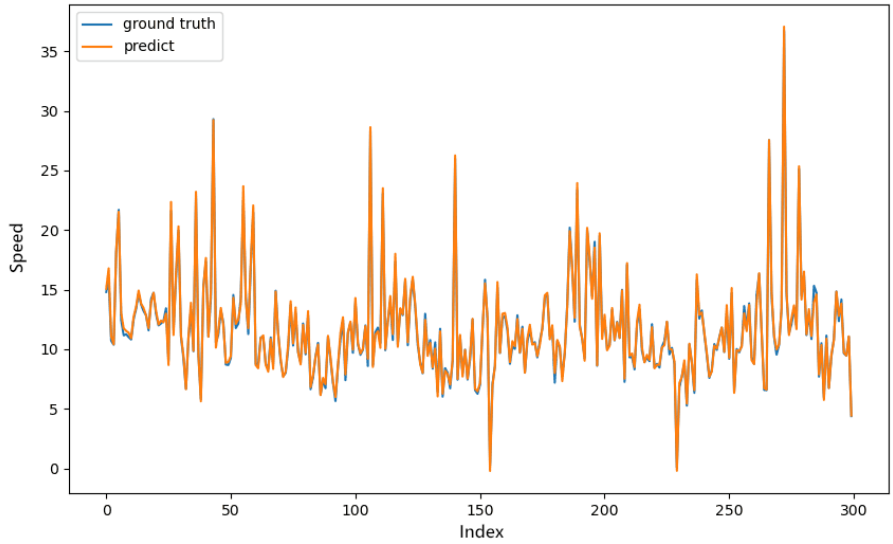


Figure11 Speed prediction based on LSTM

5.2 Traffic Flow prediction model based on LSTM

5.2.1 Data preprocessing

In 5.1, this paper first divides the taxi data in New York City in January 2020 according to CPM algorithm, and then forecasts the speed of community1 in the divided Brooklyn borough (The area labeled 217,97,66,34,256). In order to ensure uniformity, this community is still selected for traffic prediction.Since there is no direct traffic data in the original data (PULocationID(PUID), DOLocationID(DOID), LPEP_pickup_datetime (PUT), lpep_dropoff_datetime(DOT), etc.), So this paper first calculates the traffic flow.

Since this paper studies the overall traffic state of a community, it predicts the overall speed and flow of the whole community. Therefore, community 1 as a whole is regarded as a section to calculate the flow. This paper assumes that the flow of community 1 at time t is the total number of vehicles flowing in community 1 within :(t, t+10) min . Finally,the flow tables of January 2020 are obtained at 30min intervals.

5.2.2Traffic Flow prediction based on LSTM

In this paper, the time flow chart of January 2020 of community 1 is applied to the training process of the prediction model. The data are processed and input into the LSTM model training. After that, the flow in January 2020 is predicted, and the data from 2020.1.1-2020.1.10 are used to make the comparison between the predicted value and the real value, as shown in Figure11. It can be seen intuitively from the figure that the prediction model can accurately predict the evolution trend of road section traffic flow, and the predicted value is very close to the actual road section speed value at the corresponding time.Figure 12 The comparison of predicted and true valuesIn addition, it is obvious that the daily traffic flow shows a

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

very similar pattern, that is, a certain periodicity of time, which further verifies the suitability of LSTM model in traffic prediction. It also proves the feasibility of community traffic prediction

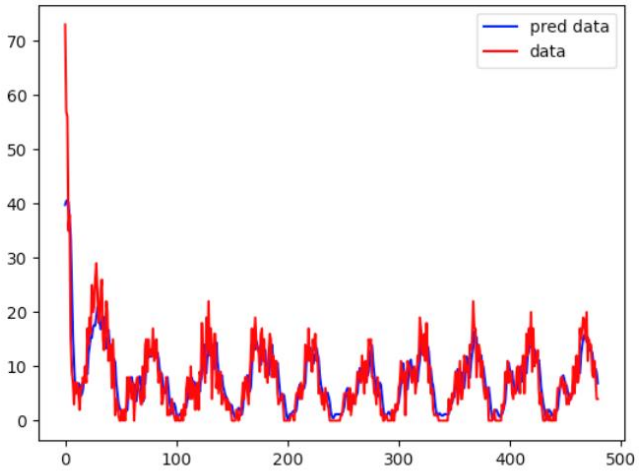


Figure 12 The comparison of predicted and true values

5.3 Fuzzy C-means clustering to classify traffic states

5.3.1 Parameter setting of FCM algorithm

In this paper, speed and flow were taken as the parameters of state discrimination, and fuzzy C-mean clustering was carried out according to the speed and flow data predicted in 5.1 and 5.2.

Firstly, this paper determines the optimal number of clustering and calculates the sum of squares of error (SSE) according to formula (13).

$$SSE = \sum_{i=1}^{j} \sum_{x \in c_i} (x - m_i)^2 \tag{13}$$

Where m is the fuzzy index and c is the number of clusters

The diagram is shown in Figure 13

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

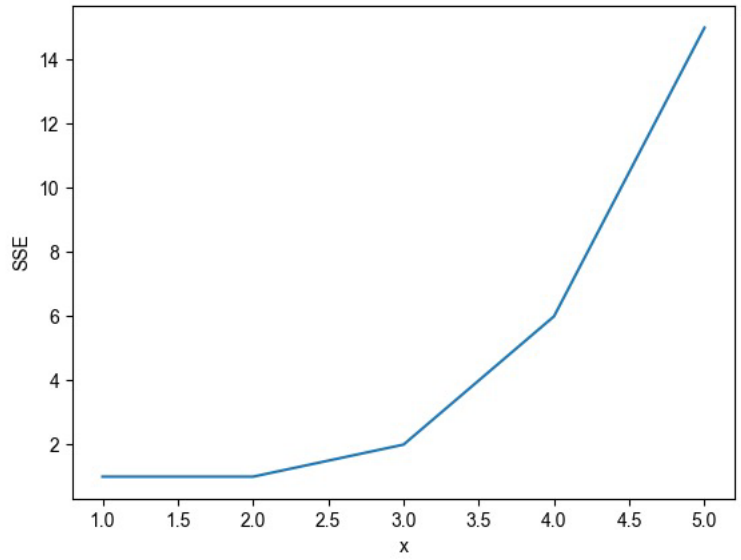


Figure 13 SSE

As can be seen from the figure, when x reaches 3, the sum of squares of error (SSE) begins to increase sharply, that is, when x exceeds 3, the error increases sharply. Therefore, the number of clustering c is selected as 3 in this paper, which is divided into three states: Open, Basically open, and mild congestion.

Other parameter Settings of FCM algorithm are shown in Table 7.

TABLE 4 Parameter setting of FCM algorithm

parameter	The values
Fuzzy index m	2
Iteration stop error ε	$1e^{-5}$
Maximum number of iterations max_iter	150

5.3.2 Analysis of classification results

This paper still uses Python to implement the algorithm.

As there may be some errors in the algorithm for obtaining flow data, a few points in the flow data deviate significantly from most of the normal points,whose velocity and flow relationship are obviously inconsistent with common sense,for example, the flow is particularly large when the speed is high, etc. After processing them, this paper conducts clustering, and the clustering results are shown in Figure 14:

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

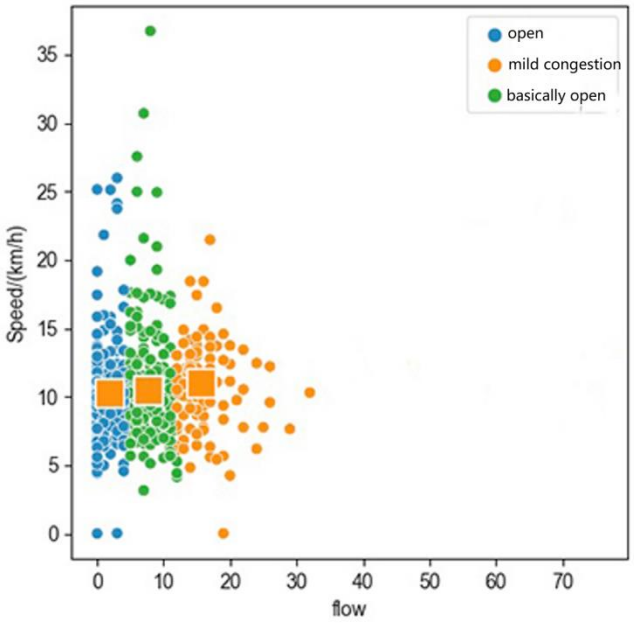


Figure 14 Clustering results

The characteristics of various states can be seen from the figure:

Open: Smooth operation of vehicles is the most expected operation state of traffic travelers. At this time, the road network flow is small and the cumulative number of vehicles is small, so drivers can maintain the expected operation state.However, at this time, the road network utilization rate is low, the road network still has a large utilization space, the flow can still increase.

Basically Open: Compared with Open, the number of accumulated vehicles in the area is increased, and the completed flow still has a trend of continuous increase, and the vehicle operation is relatively good.In the latter half, the road network runs with the highest efficiency and is the best state in traffic management.

Mild congestion: the left half is dense, and the speed of most points does not decrease significantly despite the increase of traffic flow. This shows that, at first, the road network operates smoothly and efficiently,but after that, the plate dispersion significantly increased, and the speed of most points significantly decreased, which indicates that the vehicle completion rate continued to decline with the increase of accumulated vehicles in the road network. It can be inferred from some points with the speed close to 0 that vehicles gradually appeared queuing phenomenon, and the stop-and-go. After that, the flow increases slowly until there is no increase and finally there is a traffic jam.

To sum up, this paper completed the overall traffic state prediction of community 1, and the prediction results are consistent with the actual phenomenon, proving the effectiveness of our prediction.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

VI. CONCLUSION

Based CPM algorithm,LSTM neural network model and FCM algorithm, this paper proposes a community traffic state prediction model, which predicts the overall speed and flow of the community and divides it into some traffic states. Instead of predicting a road area, it predicts a group of road areas with traffic similarity (a community),which makes it easier and more accurate to find the key features,.At the same time, compared with the prediction of single road area, the prediction of the overall traffic state of the community plays a great role in alleviating urban traffic congestion, avoiding congested communities by navigation through map app and choosing unblocked routes, and allocating orders of ride-hailing drivers to ride-hailing cars located in unblocked communities.

Based on the assumptions proposed in 3, this paper first divides communities : The adjacency map within each administrative region was first obtained, and then the CPM faction filtering topology was used to divide each region within each administrative region into several communities. Each community contained some areas with traffic similarity, and then the data of average speed was sorted and assigned to the community. Second, predict the community speed and flow: After processing the data, it is added into the LSTM network to train and predict the community speed. And the community traffic data is obtained by processing the data, which is added to the LSTM neural network for training, and the community traffic is predicted. Third , divide the traffic state of the community : Taking community speed and community flow as parameters, FCM algorithm is used to classify and finally predict the traffic state of the community.

The deficiencies of the experimental model and the areas to be improved are as follows:

First, the topology network structure constructed in this paper only considers the traffic flow correlation of the two adjacent regions, and ignores the traffic flow correlation of non-adjacent regions, so that the topology network constructed cannot completely cover the traffic correlation of each region. Since whether topological network structure can reflect the correlation of traffic characteristics of each region directly determines the community division result of CPM algorithm, and this paper divides areas with similar traffic characteristics into the same community to accurately predict the overall traffic situation of different communities in real scenes. Therefore, we should study how to set up a reasonable set of rules for constructing traffic feature networks.

Secondly, traffic state prediction usually takes speed, flow rate and occupancy rate as parameters. As there is not much content related to occupancy rate in the data in this paper, occupancy rate is not considered. In the future, more comprehensive data should be selected and occupancy rate added for improvement.

In short, the prediction model of overall community traffic state proposed in this paper predicts a group of road areas (a community) rather than one road area, so we can find key features more easily and accurately.At the same time, the prediction and classification of the overall traffic flow in the community has great advantages in alleviating urban traffic congestion, avoiding congested communities and choosing

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

unblocked routes through map app navigation, and allocating orders of ride-hailing drivers to ride-hailing cars located in unblocked communities, which has certain research and application value.

ACKNOWLEDGEMENTS

This work is supported by the National Innovation and Entrepreneurship Training Program for College Students X2021-867

REFERENCES

[1]. Chen Yun.Research on Short-term Traffic State Prediction Model of Urban Road based on LSTM Deep Network [D]. Fujian Institute of Technology,2018.
[2]. Wang Yan Ying, Huang yu, Big Data based traffic congestion evaluation Index analysis in Beijing [J]. Traffic and transportation system 1 process and information, 2016,16 (4) : 231-240.018.
[3]. Zhu Minghao. Analysis on the Social and economic Impact of Urban Traffic congestion [D]. Beijing Jiaotong University, 2013
[4]. Feng Hai-xia, WANG Qi, Yang Li-Cai, KOU Jun-jun, XIE Qing-min, ZHAO Jun-xue, MENG Xiang-lu, WANG Yan-feng. Effects of road traffic on urban air quality in congested Environment [J]. Journal of Shandong University (Engineering Science) : 1-7[2021-03-12]
[5]. Xie shenglong, Research and application of short-term traffic Flow dynamic prediction method for urban roads [Master's thesis]. Xi 'an: Chang 'an University, 2015.
[6]. Liu Jing-xin, TANG Jie, CAO Jin-xin. Urban Traffic Situation Assessment and Prediction Based on Machine Learning [J]. Journal of Inner Mongolia University (Natural Science Edition), 201,52(02):198-205.
[7]. KRAUSE B, VON ALTROCK C, POZYBILL M. Intelligent highway by fuzzy logic: Congestion detection and traffic control on multi-lane roads with variable road signs[C]//The 5th International Conference on Fuzzy Systems,1996,3:1832-1837.
[8]. Yang Qingfang, Ma Minghui, LIANG Shidong, et al. Expressway Traffic Status Discrimination Method Based on Toll Data [J]. Journal of south China university of technology (natural science edition), 2014,42 (12) : 51-57..
[9]. Huang Yanguo, Xu Lunhui, Kuang Xianxian. City road traffic status discrimination based on fuzzy c-means clustering [J]. Journal of chongqing jiaotong university (natural science),2015,34(02):102-107.
[10]. Georgia Chen, Xiaojun Wang. Short-term Traffic Flow Prediction Based on Multi-dimensional LSTM Model [J]. Journal of railway science and engineering,2020,17(11):2946-2952.
[11]. Ahmaed Mohamed S, Cook Allen R. Analysis of freeway traffic time-series data by using Box-Jenkins technique [J]. Transportation Research Record 722, 1979: 1-9.
[12]. Mascha Van Der Voort, Mark Dougherty, Susan Watson.Combining kohonen maps with arima time series models to forecast traffic flow[J]. Transportation Research Part C,1996, 4(5): 307-318.
[13]. Brian L Smith, Billy M Williams, R Keith Oswald.Comparison of parametric and nonparametric models for traffic flow forecasting[J]. Transportation Research Part C, 2002,10(4): 303-321.
[14]. Gary A Davis, Nancy L Nihan. Using time-series designs to estimate changes in freeway level of service, despite missing data[J]. Pergamon, 1984, 18(5-6): 431-438.
[15]. GUO Jianhua, HUANG Wei, Billy M. Williams. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification[J].Transportati on Research Part C, 2014, 43: 50-64.
[16]. YAO Jialin, ZHU Chuang. Research on road network speed prediction based on DCTPLS-PCA in data loss environment[J]. Journal of Railway Society and Engineering, 2019, 16(10): 2612-2619.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2022 Page No. 1623-1647
Article History: Received: 30 March 2022, Revised: 8 April 2022, Accepted: 15 April 2022, Publication: 30 April 2022

[17]. ZHANG Li, WU Huayu, LU Xiuying. Research on improved BP neural network algorithm based on rough sets[J]. Journal of Dalian University of Technology, 2009,49(6): 971-976.
[18]. DONG Chunjiao, SHAO Chunfu, XIONG Zhihua, et al.Short-term traffic flow prediction method of road network based on Elman neural network[J]. Journal of Transportation Systems Engineering and Information Technology, 2010, 10(1): 145-151.
[19]. LI Qiaoru, ZHAO Rong, CHEN Liang. Short-term traffic flow prediction model based on SVM and adaptive spatiotemporal data fusion[J. Journal of Beijing University of Technology, 2015 41(4): 597-602.
[20]. ZHAO Yaping, ZHANG Hesheng, ZHOU Zhuonan, et al.Traffic flow prediction model based on least squares support vector machine[J]. Journal of Beijing Jiaotong University, 2011, 35(2): 114-117, 136.
[21]. TAN Juan, WANG Shengchun. Research on traffic congestion prediction model based on deep learning[J].Application Research of C omputers, 2015, 32(10): 2951-2954.
[22]. LUO Xianglong, JIAO Qinqin, NIU Liyao, et al. Short-term traffic flow prediction based on deep learning[J].Applied Computer Research, 2017, 34(1): 91-93, 97.
[23]. Nicholas G Polson, Vadim O Sokolov. Deep learning for short-term traffic flow prediction[J]. Transportation Research Part C, 2017, 19: 1-17.
[24]. Hochreiter S, Schmidhuber J. Long short-term memory[J].Neural Computation, 1997, 9(8): 1735-1780.
[25]. MA Xiaolei, TAO Zhimin, WANG Yinhai, et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data[J].Transportation Research Part C, 2015, 54: 187-197.
[26]. Palla G., Derényi I., Farkas I., et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435: 814-818.
[27]. Derényi I., Palla G., Vicsek T. Clique percolation in random networks[J]. Physical Review Letters, 2005, 94(16): 160-202.
[28]. Adamcsek B., Palla G., Farkas I. J., et al. CFinder: locating cliques and overlapping modules in biological networks[J]. Bioinformatics, 2006, 22(8): 1021-1023.
[29]. Lou Ting. Research on Road Speed Prediction Model Based on Deep Learning [D]. Zhejiang University of Technology,2019.
[30]. Fan JUNpin, Li Qi, Zhu YAjie, et al. Spatio-temporal prediction model of air pollution based on RNN [J]. Science of surveying and mapping, 2017,42 (07) : 80-87+124.
[31]. Fu Chengcheng, Qin Yujun, Tian Tian, et al. An OUTBREAK prediction model for social messages based on RNN [J]. Journal of Software, 2017 (11) : 3030-3042.
[32]. Yang Shan, FAN Bo, XIE Lei, et al. Speech driven realistic face animation Synthesis based on BLSTM-RNN [J]. Journal of Tsinghua University: Science & Technology, 2017 (3) : 250-256
[33]. Lu Xian. Research on Traffic congestion Prediction Method based on Data Fusion [D]. Nanjing University of Science and Technology,2019