Accepted Manuscript

International Journal of Pattern Recognition and Artificial Intelligence

Article Title:	Application of a traffic flow prediction model based on neural network in intelligent vehicle management
Author(s):	Yang Guo, Lu Lu
DOI:	10.1142/S0218001419590092
Received:	30 May 2018
Accepted:	15 July 2018
To be cited as:	Yang Guo, Lu Lu, Application of a traffic flow prediction model based on neural network in intelligent vehicle management, *International Journal of Pattern Recognition and Artificial Intelligence*, doi: 10.1142/S0218001419590092
Link to final version:	https://doi.org/10.1142/S0218001419590092

World Scientific
www.worldscientific.com

Application of data mining technology in vehicle intelligent management system

Yang Guo[1], Lu Lu[1,2*]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510000,
People's Republic of China
[2] Modern industrial technology research institute, South China University of Technology, Zhongshan, 528400,
People's Republic of China
*lul@scut.edu.cn

Abstract: The application of intelligent technology in vehicle management system in traffic system can effectively improve its operation efficiency and security. In particular, the system is constructed under the data mining technology, and it has a good improvement on its performance. The application of data mining technology in vehicle intelligent management system mainly includes road traffic accident judgement, traffic violation research, traffic jam point judgement, parking berth judgement and so on. Taking traffic jam research as an example, this paper gives relevant algorithm formula through traffic flow prediction algorithm, traffic flow clustering analysis and traffic flow congestion event mining algorithm, improves the algorithm of blocking point, and carries out experimental analysis through samples. The experimental results show that the application of data mining technology in intelligent vehicle management system can play a good role in controlling diversion and preventing congestion, so as to improve the performance of vehicle management and promote the stable and safe development of China's transportation industry.

Keywords: data mining, vehicle management system, traffic flow, prediction

1. Introduction

With the development of economy and the increase of people's income, more and more people choose to drive a car instead of walking or cycling, causing more traffic jam which adds the pressure for traffic management. Thus, traffic management has become one of the most important duty for relating department. Modern intelligent revolution must be taken in the traditional management system which no longer suits the current situation. For example, remote sensing technique, network communication technology, data mining technology and computer technology can be combined to decrease the traffic jam, making the road safe and efficient [1]. As the core technology in all these techniques and the support for vehicle management system, data mining serves to dig useful information from lots of data.

What data mining does is to filter useful variables and their relations from lots of uncomplete, vague and random data. In practice [2], data mining is often used to describe and forecast information. It forecasts mainly the future value of important variables with the information in data sets while it describes the data pattern explained by people. From this aspect, data mining can be divided into two categories: descriptive mining and predictive mining. The process of data mining is complex, including data preparation, question definition, data conversion, data management and pattern evaluation [3]. Figure 1 shows the simple processing pattern of data mining. From the figure, it is concluded that all the steps in the pattern should be arranged in a certain order. However, the process is not linear but is reduplicate to achieve better results.

Common data mining technology often used includes association rules, clustering rules, classifying rules and so on. The use of these techniques in the intelligent vehicle system will bring data developing trend covered deeper in data which means more accurate prediction and information and more efficient traffics.

Example 1.1. Data mining technology is developing continuously. In 2004, Google first launched the non-relational data management technology represented by MapReduce. As a parallel computing model for big data analysis and processing, it soon

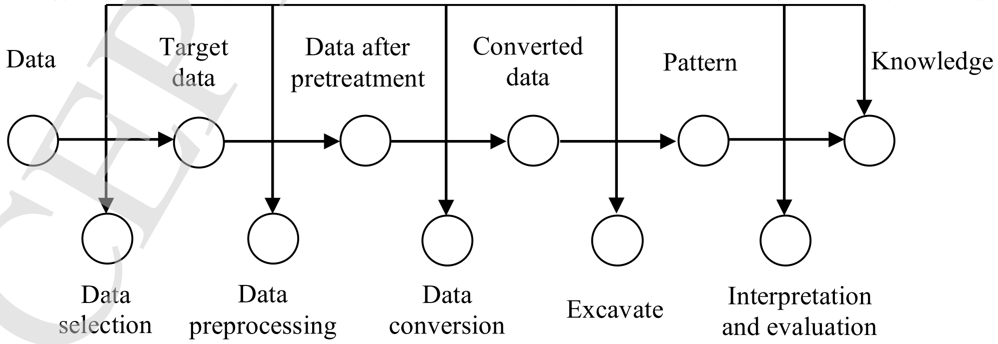

Fig.1. *Data mining process*

attracted wide attention from academia [4]. In the field of large data mining technology, scholars at home and abroad have also carried out some preliminary exploration. For example, IBM is committed to the integration of R and Hadoop for the poor extensibility of traditional analytical software and the weakness of Hadoop analysis capabilities. R is an open source statistical analysis software, which integrates the deep integration of R and Hadoop to push the computation under the parallel framework to data [5]. Another developer launched the research of Apache Mahout project, which is an open source program library based on Hadoop platform for machine learning and data mining on large scale datasets. It provides rich data analysis function for application developers [6]. In addition, aiming at data mining tasks such as frequent traffic pattern, classification and clustering, researchers also put forward corresponding big data solutions. For example, when Tatti et al. [7] studied Houston's public transport intelligent management system, MapReduce and K-Center and K-Median clustering methods under large-scale data were implemented. David et al. [8], in the research of intelligent transportation management, researches the management of large-scale graph data and the access characteristics of local graph (breadth first query and random walk) based on cluster, and puts forward the distributed graph data environment and two-level division management framework. Lars et al. [9] is faced with the dredging plan of Los Angeles's urban traffic management. It proposes an extensible algorithm for mining frequent sequential patterns in MapReduce framework. Fu Xiao et al. [10] introduces the large data classification method for linear classification model into urban intelligent traffic management and has achieved some results.

Example 1.2. A notable feature of big data is that data volume accumulates rapidly with time. People can make full use of historical data and new data to analyze the state of objects and predict the trend of events [11]. The data mining tasks in many practical applications have high timeliness requirements. For example, data from Shanghai traffic administration's monitoring platform are dynamically increasing and updating all the time. Decision makers need to grasp the behavior mode and road traffic trend in a timely manner, so as to dredge traffic congestion more accurately [12]. In the vehicle intelligent management system of Guangzhou, how to accumulate real-time vehicle information and how to make timely optimization and dredging decisions is very important for reducing road congestion and improving road patency [13]. In the research field of vehicle intelligent management, the state of a large number of nodes is constantly changing, and the link between nodes and nodes is constantly changing. This brings challenges to the real-time requirement of data mining of vehicle intelligent management [14]. OLAP (online analytical processing) is put forward to fit the timeliness requirements of this data analysis. However, in the era of big data, data growth is not only reflected in the fast, but also reflects the mass of data required by the incremental time period in the analysis, the limitations of the traditional single or small batch incremental machine learning technology highlights, online data big data calls for more efficient analysis techniques.

Telecommunication data mining has always been the most urgent problem for mobile service providers. In the past, Providers would often base their solutions on integrated mining, which limited their performance in data subdivision. What is innovative in the paper is that it data mining for mobile information combining rough set and neural network analyses the changing law of mobile nodes from both moving direction and moving distance. It uses two-layer neural network with nonlinear connection weights as the information distinguishing system, simplifying the network learning and training samples with the method of rough set. Based on the mobile information node changing laws, the data mining generalizes input and output data, deleting redundant data and attributes in the input information table and calculating the simplified attribute set, resulting in the simplest table. Later, it learns and trains with the simplified sample set until the ideal mining accuracy is achieved. We use the finished neural network to calculate, distinguish and classify the mined data and find the result. Simulation shows that this method reflects high mining accuracy and efficiency.

2. Application of data mining in vehicle intelligent management

The application of data mining technology of vehicle intelligent management is mainly to collect, statistics and analyze the following types of data. It includes road traffic accident study, traffic law study, traffic jam study, parking research and so on.

2.1. Analysis of road traffic accidents

(1) Hot spot analysis

According to the area, time, type, cause and so on, we analyze the road traffic accidents, find out the locations of all kinds of accidents and stack the distribution of police force through data mining, and analyze the rationality and effectiveness of the distribution of police forces.

(2) Polymerization analysis

According to the area, time, type, reason and so on, we can aggregate the road traffic accidents. We can count the number of traffic accidents in various accidents by data mining, and we can superimpose the layers of police force distribution and analyze the rationality and effectiveness of the police force distribution.

(3) 24-hour analysis

According to the types of accidents, accident form, accident consequences, the person responsible for the accident, driving the person responsible for the accident, accident type, driving road area of 24 hours on the road traffic accident analysis, mining of relevant data, analysis of the change regularity of various types of road traffic accidents in a day, find out all kinds of road traffic accidents in time.

(4) The analysis of the week date

According to the types of accidents, accident form, accident consequences, the person responsible for the accident, driving the person responsible for the accident, accident type, driving road area on the road traffic accident information mining, the implementation date of week analysis, analysis of the change regularity of various types of road traffic accidents within a week, Find out the high day of all kinds of road traffic accidents.

2.2. Analysis of traffic law

(1) Hot spot analysis

According to the time, type, cause and so on, the hotspots of traffic violations are analyzed. Through data digger, we can find out multiple locations of illegal behaviors and stack up the distribution of police force, so as to analyze the rationality and effectiveness of police force distribution.

(2) Polymerization analysis

According to the time, types and reasons, we can aggregate the road traffic accidents. We can count all kinds of illegal numbers of traffic violations by data mining. We can superimpose the layers of police force distribution and analyze the rationality and effectiveness of the police force distribution.

(3) 24-hour analysis

According to the results, the illegal punishment type, vehicle type, responsibility, responsibility for driving 24 hours of driving, analysis of traffic violations through data mining and analysis of all kinds of traffic violations within a day of the changes, find out all kinds of traffic violations in time.

(4) The analysis of the week date

According to the results, the illegal punishment type, vehicle type, responsibility, responsibility for driving, week date of traffic violations through the analysis of data mining, analysis of the change regularity of various types of traffic violations within a week, find out all kinds of traffic violations in japan.

2.3. Traffic jam judgement

(1) Hot spot analysis

According to the time, types and causes of the traffic blocking point of focus, through the data mining to identify congestion prone, superposition police distribution, distribution facilities, green wave signal distribution, bus lane distribution layer, analysis of the police, signal facilities, green belt, bus lanes and the distribution of rationality and validity.

(2) Polymerization analysis

According to the time, types and causes of the traffic blocking point of focus, through the data mining to find the number of congestion prone to congestion, superposition police distribution, distribution facilities, green wave signal distribution, bus lane distribution layer, analysis of the police, signal facilities, green belt, bus lanes distribution is reasonable and effective of.

(3) 24-hour analysis

According to the area, reasons, etc., through data mining, we analyze the traffic jam data for 24 hours, analyze the change rule of traffic jam in one day, and find out the high occurrence time of all kinds of traffic jams.

(4) The analysis of the week date

According to the area, reasons and so on, the date of traffic jam data is carried out. Data mining is used to analyze the variation rule of traffic congestion in a week and find out all kinds of traffic congestion high incidence days.

2.4 Parking research

(1) 24-hour analysis

Through data mining, the parking berth is analyzed for 24 hours, and the peak time of parking peak is found.

(2) The analysis of the week date

The week date of parking berth is analyzed by data mining, and the peak day of parking is found.

3. The Algorithm and Model of The Application of Data Mining Technology in The Intelligent Vehicle Management System

3.1. Traffic flow prediction algorithm

As early as the late 80s of last century, many foreign researchers have put forward many short-term traffic flow prediction models, from historical average model to highway traffic flow prediction method based on ARIMA model and ARIMA model applied to single point traffic flow prediction [15]. However, because of the high demand for data by ARIMA model, it can only satisfy the deterministic traffic flow prediction, because the exception caused by ARIMA model has great influence on the model. And the ARIMA model is an off-line process that can not be adjusted adaptively. Meanwhile, Calman filtering theory is also built by user prediction model, but it is a linear model, and its performance is low when traffic flow is uncertain or nonlinear. In 1991, western traffic experts put forward a nonparametric regression model. In 2002, a parameter regression model was put forward. Then a neural network model based on time delay was put forward [16]. For traffic flow prediction, we design a CMTFPA Combined Model Traffic Flow Prediction Algorithm, which combines traffic flow sequence segmentation with neural network. The algorithm first divides traffic patterns in traffic and time, and then uses neural network to predict each traffic flow. This method has a good prediction precision for a single neural network model.

3.1.1 Traffic flow sequence segmentation: In recent years, traffic flow data as a research hotspot of time series analysis, is based on time series data change segmentation, in fact, it is a form of time series [17]. The most commonly used method is to divide the sequence with a linear model and describe one by one, that is, piecewise linear description [18]. In this paper, traffic flow is divided on the basis of traffic clustering method. So, the K-Means clustering algorithm is selected as the basic algorithm, and the basic idea is as follows:

(1) According to the traffic flow value of different sizes, the K-Means algorithm is used. set up k＝7, The distance function uses Euclidean distance;
(2) Three step cluster results S1,S2,S3. Hypothesis S1＞S2＞S3, The first cluster is aggregated into two clusters by clustering algorithm again, S2 is aggregated into three clusters, and S3 is polymerized into two clusters. It needs to be explained that the clustering of this step is based on the size of the time, not according to the traffic flow value. In this step, S1 and S2 are assembled into five clusters.
(3) Through the above two steps, we will aggregate the daily road traffic volume into seven clusters, and some clusters and clusters will have data overlap. However, according to the time conditions, we can query the order generated successively.
(4) Next, we calculate the centroid and mass of each cluster. The time coordinate is placed on the abscissa of the center of mass. The ordinate shows the average value of the flow. The number of objects in the cluster is the quality of the cluster. According to the coordinates of the time, the size of the time is set to T1, T2, T3, T4, T5, T6, T7, Quality Q1, Q2, Q3, Q4, Q5, Q6, Q7.
(5) Calculate the time according to the following formula T_1^*, T_2^*, T_3^*, T_4^*, T_5^*, T_6^*, T_7^* value.

$$T_1^* = T_1 + \left|\frac{Q_1}{Q_{2-}Q_1}\right| \times (T_2 - T_1) \qquad (1)$$

In the above algorithm, according to the two dimensions of vehicle flow and time, clustering algorithm is used to divide traffic flow patterns, which is fully consistent with the actual traffic flow distribution. There are two peaks in traffic flow: early peak and late peak. We want to divide the flow of the peak and classify the flow with the value of the volume. For a city's main intersection, a traffic package contains huge traffic information, and the traffic command scheme is not fixed. It must be changed according to the real-time traffic condition at that time. Many methods of adjustment programs, can be adjusted according to the fixed time, also can adjust to the main road green as the main control points to adjust [19]. All of the above adjustments can be mapped out by the change of the traffic flow value of the road, and the two-peak data can also be divided in the time dimension clustering division.

3.1.2 BP neural network prediction model: The BP neural network is also called the error back propagation network, and the structure is shown in figure 2. This network is essentially a front - to - no feedback network. The structure is clear, the computing function is very powerful, and it is easy to implement [20].

The BP neural network is divided into four layers, one input layer and second output layer, containing one or more hidden layers [21]. Each layer has a node with an indefinite number of modules, each node is a neuron, and there is no connection between the nodes on the same layer. Between each layer and each layer, there is a full connection between each neuron. Following the supervised learning mode, once the information is transmitted to the network, it will stimulate the neurons. The activated values first go through the input layer, then pass through each hidden layer node, and the final network inputs each node that is mapped to the output layer. In the aspect of traffic forecasting, the neural network has achieved certain effect, so BP neural network is chosen as the prediction model of flow prediction. The characteristic of this model is that every layer of neurons is connected to the adjacent layer only. There is no connection between neurons in each layer. There is no feedback connection between neurons in each layer. The three-layer BP network learning process mainly has the following four parts:
(1) the input mode is propagating: the function of this process is mainly to use the input mode to solve the corresponding actual output. The set input vector is X_k, then there is:

$$X_k = x_1^k, x_2^k, \dots\dots x_n^k \qquad (2)$$

In the form, k = 1,2,⋯⋯n, is the logarithm of the learning pattern (or the logarithm of the training pattern); n is the number of elements in the input layer. Set the desired output vector to be Y_k, the formula is:

$$Y_k = y_1^k, y_2^k, \dots\dots y_q^k \qquad (3)$$

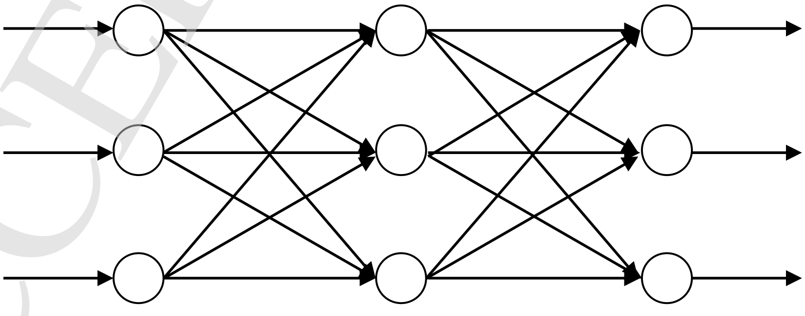

Fig.2. *BP neural network structure diagram*

In the form, $k = 1,2, \ldots \ldots q$ is the logarithm of the learning pattern (or the logarithm of the training pattern); q number of units for the output layer. Then the activation value of each neuron in the hidden layer is S_j, The formula is:

$$S_j = \sum_{i=1}^{n} w_{ij}\, x_i - \theta_j \qquad (4)$$

In the form, n is the number of elements in the input layer; w_{ij} is the connection weight value (the input layer to the hidden layer); the θ_j is the threshold (implicit layer unit); and $j = 1,2, \ldots \ldots p, p$ is the unit number of the hidden layer. So, the activation function uses the *S* type function, the formula is:

$$f(x) = \frac{1}{1 - \exp(-x)} \qquad (5)$$

The output value of the *j* cell in the hidden layer is calculated. The activation value of the above formula, that is, the formula (4), is added to the activation function formula (5), and the output value of the hidden layer is calculated.

$$b_j = f(S_j) = \frac{1}{1 - \exp\left(-\sum_{i=1}^{n} w_{ij}\, x_i\right) + \theta_j} \qquad (6)$$

In the above formula, during the learning process, the threshold value theta θ_j is constantly updated and updated, which is similar to the weight value w_{ij}. Then the activation value of the *t* unit of the output layer is calculated: O_t, the formula is:

$$O_t = \sum_{j=1}^{p} \left(w_{ij} x_j\right) - \theta_t \qquad (7)$$

And the actual output value of the *t* unit of the output layer is C_t, the formula is:

$$C_t = f(O_t) \qquad (8)$$

In the above formula, w_{ij} is the weight of the hidden layer to be transmitted to the output layer; the θ_t is the unit threshold of the output layer; of which, $j = 1,2, \ldots \ldots p, p$ is the unit number of the hidden layer; x_j is the output value of the *j* node in the hidden layer; *f* is a S type activation function, of which $t = 1,2, \ldots \ldots q$, and q is the number of units in the output layer. The above formulas can be used to calculate the CIS propagation process of an input mode.

(2) the inverse propagation of the output error: in the first step of the pattern CIS propagation calculation, the actual output value of the network is obtained. When these data have errors or are not consistent with the expectations of the situation, it is necessary to correct the network. The correction is carried out in the past, so it is called the error inverse propagation. Its computing process is to transfer from the output layer to the hidden layer and then transfer from the hidden layer to the input layer. The formula for the correction error of the output layer is as follows:

$$d_t^k = (y_t^k - c_t^k) f'(\mathrm{o}_t^k) \qquad (9)$$

In the above formula, $t = 1,2, \ldots \ldots q$, q is the number of units of the output layer; in which, $k = 1,2, \ldots \ldots m$, m is the logarithm of the training (or learning) pattern; y_t^k is the expected output value; c_t^k is the actual output; $f'(\mathrm{o}_t^k)$ is the derivative of the output function. Thus, the correction error formula for each unit of the hidden layer is as follows:

$$e_j^k = \left(\sum_{t=1}^{q} w_{jt} d_t^k\right) f'(\mathrm{s}_t^k) \qquad (10)$$

In the formula, $t = 1,2, \ldots \ldots q$, q is the number of units of the output layer; $j = 1,2, \ldots \ldots p$, p is the unit number of the hidden layer; $k = 1,2, \ldots \ldots m$, m is the logarithm of the pattern of training (or learning). The threshold correction formula for the connection weight and the output layer of the output layer to the hidden layer is as follows:

$$\Delta v_{jt} = a d_t^k b_j^k \qquad (11)$$

$$\Delta \gamma_t = a d_t^k \qquad (12)$$

In the formula, b_j^k is the output of the hidden layer J unit; d_t^k is the correction error value of the output layer, in which $j = 1,2, \ldots \ldots p$; $t = 1,2, \ldots \ldots q$; $k = 1,2, \ldots \ldots m$; a＞0 (the output layer is transmitted to the learning rate of the hidden layer). The formula for the correction of the hidden layer to be transmitted to the input layer is:

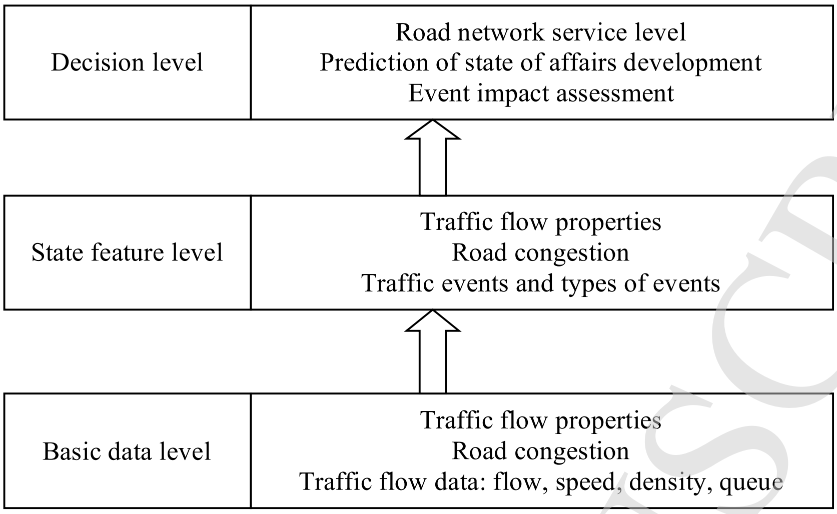
Decision level
Road network service level
Prediction of state of affairs development
Event impact assessment
State feature level
Traffic flow properties
Road congestion
Traffic events and types of events
Basic data level
Traffic flow properties
Road congestion
Traffic flow data: flow, speed, density, queue

Fig.3. *Traffic information stratification graph*

$$\Delta \mathrm{w}_{ij} = \beta e_t^k x_i^k \qquad (13)$$

$$\Delta \theta_j = \beta e_t^k \qquad (14)$$

In the formula, e_t^k is the correction error value of the j element in the hidden layer; x_i^k is the standard input value, in which $i = 1,2, \dots \dots n$, *n* is the number of elements in the input layer; of which $0<\beta<1$ (implicit layer to the learning rate of the input layer).
(3) cyclic memory training (learning): in order to ensure the output network error rate reaches a minimum, so BP neural network to the input of each group learning mode, to go through tens of thousands of times or with high frequency cyclic memory training (or learning), which can promote the network to remember this group mode. This kind of circular memory is actually the calculation of the first and third steps.
(4) Discriminative training results: cyclic memory each time after the end of the training, will be on the training results for discrimination, its main purpose is to test the output error is the smallest in the allowable range, if the result is in the allowable range can finish the training process, on the other hand, the memory training cycle continue.

3.1.3 Traffic flow forecasting combined model: BP neural network because of its advantages of simple structure, stable, easy to realize hardware is widely used, but there are also some shortcomings, specific as follows: neural network model is composed of a plurality of groups of piecewise function composed of lead to the input and output data of the prior knowledge is not easy to integrate into the model that influence the accuracy of recognition; secondly, in the training process, need a large original clean data, is not easy to build model. Therefore, there is a certain distance between the BP neural network method and the practical engineering application. In traffic volume prediction, once the network facilities and traffic conditions of the road change, the network that is built must be updated timely, so that we can continue to analyze and predict.

The neural network structure is complex and the network structure is simple, once the time is long, it is difficult to solve the problem in real time prediction. Because intersection is different from other road traffic in complexity and particularity, so if BP neural network works alone, it needs to predict its short-term traffic flow and improve its accuracy. Therefore, we have designed a multi model algorithm to improve the accuracy of the prediction. The central idea is based on traffic sequence segmentation algorithm, and neural network (BP) is used to do prediction algorithm. The algorithm combines the algorithm to model and predict the captured data.

The above combined model algorithm is applied to predict traffic flow, and then the algorithm of successive clustering is carried out on two dimensions of traffic and time, and then traffic patterns are divided to conform to traffic distribution. From the above analysis, it is very easy to see that the flow shows two peaks, an early peak and a late peak. And we want to divide the flow of traffic at peak time, we need to cluster the flow with the value of the volume. For a city's main intersection, a traffic package contains huge traffic information, and the traffic command scheme is not fixed. It should be changed according to the real-time traffic conditions at that time. Many methods of adjustment programs, can be adjusted according to the fixed time, also can adjust to the main road green as the main control points to adjust. All of the above adjustments can be mapped out by the change of the traffic flow value of the road, and the two-peak data can also be divided in the time dimension clustering division.

Practice has proved that in the rush hour period, such as the morning hours, the two predictions have little difference, while in other stages, the error rate of the prediction algorithm using the combination model is greatly reduced, which can provide effective information for traffic management and prediction.

3.2. Traffic flow clustering analysis

Generally speaking, road traffic flow is described by popular words (smooth and crowded), quantitative indicators (service level) and so on, so as to map the stage and level of the development and change of the system. As shown in figure 3, it is a hierarchical network traffic information map. It can be seen from the low to high score: basic data, state characteristics and three levels of decision-making. The lowest level data is usually detected by various kinds of detection devices, such as traffic and speed. It is the foundation of upper level data. The second level is traffic flow status, events and types. The first level describes information based on these display decisions, evaluation and prediction. This is also a process of data mining.

The processing of the bottom data is analyzed, and the characteristics of traffic flow state are analyzed. For single traffic attributes, such as speed, process and queuing example, it can not reflect the state characteristics of the current traffic flow. Earlier, some traffic flow models were taken as the object of study. A quantitative identification index for traffic situation based on traffic flow, road static characteristics, vehicle speed and events were proposed. However, for the traffic flow model, it is usually reasonable to simplify the research question, and static index, we constructed the physical model of the actual situation of the road cannot be completely taken into account, so that in the process of the fuzzy judgment, so as to achieve good results.

3.3. Traffic flow congestion event mining algorithm

In the original traffic flow data, in a similar search matching using Euclidean distance metric, Euclidean distance is defined as follows: two mode PA_1, PA_2 speed were V_1, V_2, flow were F_1, F_2, or share were or_1, or_2 w is set to the weight of weight. Then the distance formula between the two modes is:

$$D(PA_1, PA_2) = \sqrt{w_{1(v_1-v_2)^2}+w_{2(f_1-f_2)^2}+w_{3(or_1-or_2)^2}} \quad (15)$$

During the test, all the data in the traffic flow data and the model library should be displayed in layers according to the algorithm. After the preprocessing is completed, the algorithm searches for the matching. Each group of data contains three variables: V, F and OR. Each item is matched with the original data of the schema library. Then, in order to facilitate the work, we perform linear synthesis according to certain weights after matching three variables. Finally, the data of the current time and the historical data in the schema library are similar to each record, and finally, the correct judgment is made. When new patterns occur in the road situation, the simple way to update the schema library is to add it to the schema library.

With the rapid increase of private cars in today's cities, the original traffic data are becoming more and more unable to adapt to the rapid development of traffic information. Therefore, we should consider the actual application of data mining. We need to find a suitable method to make prompt judgement in the second level time. When updating the mode library, we should use the deletion strategy to update the schema base, while maintaining the consistency of the schema library.

As a result, the above algorithm is optimized. First, when a pattern database is created, the traffic will increase with the increasing over time, you must update the pattern library; second, first model library determines the size of the matching algorithm in time, therefore, to reduce the matching time, that is to say, in the shortest time, reduce the algorithm is time. To optimize the pattern library. According to this optimization idea, the improved mining algorithm is given below, as shown in Table 1.

Table 1 Classification of end users

Classification of end users
Input: current time t and traffic flow data TF; Output: t time traffic condition TC. Method： int n = number(S); Obtain $V_{t,}$ $F_{t,}$ O, R_t
according to $V_{t,}$ $F_{t,}$ O, R_t Predefined values are first divided into TF
Using the algorithm to get the current situation of the traffic road situation S

if(M) // If S is not empty
return M;
else{
call Bulid-C(TF, t, t+1);}
for(int k=O; k<n; k++){

Delete the corresponding record when the least state is deleted;

Return M;
}

4. Experiment Result

According to the improved mining algorithm, samples are prepared. Sample preparation: In sample preparation process, traffics in 15 minutes after time twill first be forecasted. The identification of the forecasted target section is [link id, start.(11, nodeid end node id)6021 6004]means that the forecasted samples are saved in the traffic data of all the time sections between

21th October, 2017 and 27th October, 2017. Select the sections which meet the rule that start-node-id is 6004 while end-node-id is 6021 using T_DIRECTIONAL_LINK.

```
select link-id, start-node-id, end-node-id from T-DIRECTIONAL-LINK
where start-node-id=6004
or end-node-id=6021;
```

There are six sections which meet our demand, in which (8,6004,6005), (29,6004,6007), (12,6004,6023) are downstream sections while (14,6024,6021), (32,6025,6021), (33,6027,6021) are upstream sections. The traffics and vehicle speed information in these sections including those of the target sections between 21st October, 2017 and 27th October as well as the traffic information of targeted sections in the next time section is the largest sample in the training. Next, we classify the samples according to designed algorithm. First is the relevant K. We first calculate the maximum vehicle speed of the three sections (8,6004,6005), (29,6004,6007), (12,6004,6023) in the sample. For example, for (8,6004,6005)：

```
select max(spoLspd) from T-FLOW-MONTH
where start-node-id = 6004 and end-node-id = 6005
and check-date between to_date('2006-9-1', 'YYYY-MM-DD')
and to_date('2006-9-7', 'YYYY-MM-DD');
```

The maximum speed in the section can be up to 49.671. Then the maximum of the other two sections are also calculated. As shown in Table 2.

Lee et al. [22] tried to use rough set to analyze the objective factors of road traffic accidents and calculate the maximum speed of each node of the traffic jam. But the research they carried out did not classify the sections upstream and downstream, and the maximum speed of the vehicle was only a uniform speed. Compared with their research, the study of this paper is more detailed and accurate for the calculation of the maximum speed of vehicles in each section.

Through the research of intelligent data mining, the experiment of traffic flow prediction algorithm, BP neural network forecasting model, traffic congestion, traffic flow prediction model of combined mining is analyzed, and stream clustering and traffic on the traffic flow jam mining algorithm is analyzed, the mining algorithm was given improved. The traffic flow data showed two peaks: first, according to the characteristics of the morning and evening peak, each prediction model, so as to improve the prediction accuracy; secondly BP neural network prediction model was established for mining, traffic congestion, traffic flow prediction combination model and clustering combined with traffic flow, the data mining algorithm for design the development of the system.

The rule excavated from the data mining is that it is suggested to maintain the road at 2:30 am as the traffic flow decreases. Traffic flow reaches its peak at 7:20 -10:30 and 17:40-20:30. Notice board is advised and the information should be sent to drivers, helping them avoid the peak according to their own situation, which may effectively solve the problem of traffic jam.

Table 2 Calculation results of maximum value of 6 sections

Section name	*t*	start-node-id	end-node-id	Maximum speed
Section 1	8	6004	6005	49.671
Section 2	29	6004	6007	48.802
Section 3	12	6004	6023	50.613
Section 4	14	6024	6021	50.974
Section 5	32	6025	6021	51.386
Section 6	33	6027	6021	51.412

5. Conclusion

Due to the large population and the high density in China, traffics have always been the focus of Chinese road management. The use of data mining technology can arrange vehicles and solve the traffic jams in rush hours more effectively. The problems of card theft and license plate theft can also be solved, increasing information value, making traffic management more effective and fulfilling the safe and efficient development of our traffic management.

Traffic flow prediction algorithm based on combination model can accurately predict the traffic intersection in a short period of time, effectively improve the current traffic flow forecast, can meet the traffic administrative supervision and high efficiency of transportation enterprise management, public traffic service requirement. But the following four aspects should be improved:

(1) It is necessary to further improve the traffic flow pattern library and further study the partition mining algorithm.

(2) In traffic flow congestion mining algorithm, we study the similarity measurement algorithm of new data and raw data, and we can consider the design of multi-dimensional time series similarity to improve the search algorithm.

(3) Further accumulation and improvement of system access to traffic data information;

(4) For the system to further improve and expand the functions of the system, and improve the efficiency of the system, we study the corresponding data mining algorithm and improve the application scope of the system.

6. Acknowledgment

This research was supported by the National Nature Science Foundation of China (No. 61370103), Guangdong Province Science & Technology Fund (2015B010129007), Guangzhou Province Science & Technology Fund (201802020006) and Zhongshan Produce & Research Fund (2017A1014).

7. References

[1] Sun, Lu Pan, Yiyong Gu, Wenjun. Data mining using regularized adaptive B-splines regression with penalization for multi-regime traffic stream models. Journal of advanced transportation, 2014, 48(7), pp. 876-890.
[2] Alejandro Pena-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Application, 2014, 41(4), pp. 277-289.
[3] Alfredo Cuzzocrea, Carson Kai-Sang Leung, Richard Kyle Mackinnon. Mining constrained frequent itemsets from distributed uncertain data. Future generations computer systems: FGCS, 2014, 37(8), pp. 722-731.
[4] Gangin Lee, Unil Yun, Keun Ho Ryu. Sliding window based weighted maximal frequent pattern mining over data streams. Expert Systems with Application, 2014, 41(2), pp. 87-102.
[5] Stefan Strohmeier, Franca Piazza. Domain driven data mining in human resource management: A review of current research. Expert Systems with Application, 2013, 40(7), pp. 385-396.
[6] Shelokar, P.Quirin, A.Cordón, O. A multiobjective evolutionary programming framework for graph-based data mining. Information Sciences: An International Journal, 2013, 23(11), pp. 123-133.
[7] Tatti, N.Vreeken, J. Comparing apples and oranges: Measuring differences between exploratory data mining results. Data mining and knowledge discovery, 2012, 25(2), pp. 39-50.
[8] David John Stockton, Riham Ashley Khalil, Lawrence Manyonge Mukhongo. Cost model development using virtual manufacturing and data mining: part II - comparison of data mining algorithms. The International Journal of Advanced Manufacturing Technology, 2013, 66(9), pp. 108-121.
[9] Lars Graening, Bernhard Sendhoff. Shape mining: A holistic data mining approach for engineering design. Advanced engineering informatics, 2014, 28(2), pp. 271-285.
[10] Fu Xiao, Cheng Fan. Data mining in building automation system for improving building operational performance. Energy and buildings, 2014, 75(6), pp. 112-124.
[11] Marko Debeljak, Ales Poljanec, Bernard Zenko. Modelling forest growing stock from inventory data: A data mining approach. Ecological indicators: Integrating, monitoring, assessment and management, 2014, 41(6), pp. 218-229.
[12] De Chen, Long Chen, Jing Liu. Road link traffic speed pattern mining in probe vehicle data via soft computing techniques. Applied Soft Computing, 2013, 13(9), pp. 3894-3902.
[13] Wei He, Tao Lu, Enjun Wang. A New Method for Traffic Forecasting Based on the Data Mining Technology with Artificial Intelligent Algorithms. Research journal of applied science, engineering and technology, 2013, 5(12), pp. 3417-3422.
[14] Eugenio Cesario, Carlo Mastroianni, Domenico Talia. A Multi-Domain Architecture for Mining Frequent Items and Itemsets from Distributed Data Streams. Journal of grid computing, 2014, 12(1), pp. 98-109.
[15] Nada Lavrac, Petra Kralj Novak. Relational and Semantic Data Mining for Biomedical Research. Informatica: An International Journal of Computing and Informatics, 2013, 37(1), pp. 227-236.
[16] Mhmood Deypir, Mohammad Hadi Sadreddini, Mehran Tarahomi. An Efficient Sliding Window Based Algorithm for Adaptive Frequent Itemset Mining over Data Streams. Journal of information science and engineering: JISE, 2013, 29(5), pp. 44-57.
[17] Subbiyan Prakash, Murugasamy Vijayakumar. An Effective Network Traffic Data Control Using Improved Apriori Rule Mining. Circuits and Systems, 2016, 7(10), pp. 3162-3173.
[18] Bhardwaj Amit Kumar, Singh Maninder. Data mining-based integrated network traffic visualization framework for threat detection. Neural computing & applications, 2015, 26(1), pp. 117-130.
[19] G. Manikandan, S. Srinivasan. The Application of Spatial Data Mining in Traffic Geographic Information Systems. Advances in computational sciences and technology, 2012, 5(2), pp. 101-110.
[20] Pablo Velarde-Alvarado, Rafael Martinez-Pelaez, Joel Ruiz-Ibarra, Victor Morales-Rocha. Information Theory and Data-Mining Techniques for Network Traffic Profiling for Intrusion Detection. Journal of Computer and Communications, 2014, 2(11), pp. 24-30.
[21] Jaehak Yu, Hyunjoong Kang, DaeHeon Park, Hyo-Chan Bang, Do Wook Kang. An in-depth analysis on traffic flooding attacks detection and system using data mining techniques. Journal of systems architecture, 2013, 59(10), pp. 1005-1012.
[22] Lee, W.-H., Tseng, S.-S., Shieh, J.-L., Chen, H.-H. Discovering Traffic Bottlenecks in an Urban Network by Spatiotemporal Data Mining on Location-Based Services. IEEE transactions on intelligent transportation systems, 2011, 12(4), pp. 1047-1056.

8. Authors Information

Yang Guo, 1981, is currently a doctoral student at School of Computer Science and Engineering, South China University of Technology, P.R.China. His main research interests are software engineering, big data analysis and artificial intelligence.

Lu Lu, 1971, received his Ph.D in 1999, Xi'an Jiaotong University, now he is professor in the department of computer science, South China University of Technology, P.R.China. His main research interest is software engineering, software testing and software architecture design.