Prediction Traffic Flow with Combination Arima and PageRank

Cheng-fan Li[1,2], Jia-xin Huang[1], and Shao-chun Wu[1(✉)]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
{lchf,scwu}@shu.edu.cn, 1204833945@qq.com
[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200044, China

Abstract. Modern traffic network information is similar to the complex network structure in that the links between the sections are quite complex. Therefore, predicting the traffic flow between sections can effectively relieve traffic congestion. To solve this problem, this paper proposes a combined model of Arima and PageRank to predict the traffic flow of each section of the road network. First, the trained Armia model is used to predict the average speed and traffic flow of each section, and then the PageRank model is used to calculate the weight of each section. The product of traffic flow and weight is output as the final result. Through the experiment of highway traffic data in PeMS database, this method is verified to be able to predict the traffic flow of the whole road network.

Keywords: Network structure · Traffic congestion · Armia model

1 Introduction

Traffic congestion not only affects people's life in terms of travel, but also causes economic losses, environmental problems and safety problems [1–5]. Such problems cannot be completely solved only by improving road facilities, so intelligent transportation emerges at the right moment.

At present, there are many models proposed to solve traffic prediction, for example, based on statistical theories, there are chaos theory [6], Kalman filtering method [7] and so on; based on non-linear model, there are LSTM method [8], k-nearest neighbor [9] and so on; based on artificial intelligence method, there are neural network [10], Bayesian probabilistic neuron [11] and so on; based on combination model, there are RBF neural network, fuzzy c-mean combination model [12] and based on combination of support vector machine and data demising schemes [13].

The above methods aim at single point prediction and cannot meet the prediction of traffic flow between each section of the road network. In this paper, considering the shortcomings of the existing prediction models, a combined model of Arima and PageRank is proposed to realize the prediction of the traffic flow of the whole network. First, the combined model can predict the traffic flow of each section of the road network.

D.-J. Deng et al. (Eds.): SGIoT 2019, LNICST 324, pp. 129–140, 2020.
https://doi.org/10.1007/978-3-030-49610-4_11

Secondly, the model can calculate the weight of each section in the road network. Third, this method is beneficial to the drivers to choose the better route. Fourthly, the combined model can also calculate the arrival time between sections. When possible congestion time on the road is predicted, road operators will take corresponding measures to avoid traffic congestion.

2 ARIMA Speed and Flow Prediction Model

ARIMA (p,d,q) is a combination of Autoregressive model (AR), Moving Average model (MA) and difference method (I). In this paper, speed V and flow F are obtained by using ARIMA model. In order to reduce the article length, ARIMA formula was used to derive $\mathrm{V} = \{\mathrm{Vm}, \mathrm{m} = 1, 2 \ldots . \mathrm{m} - 1\}$. Vm is the output value and the establishment of the model requires the following three steps.

Step1: The purpose of data smoothing is to transform the non-stationary of data sequence into stationary. In order to highlight the smoothing effect, this paper directly performs visual processing of data, as shown in Fig. 1.

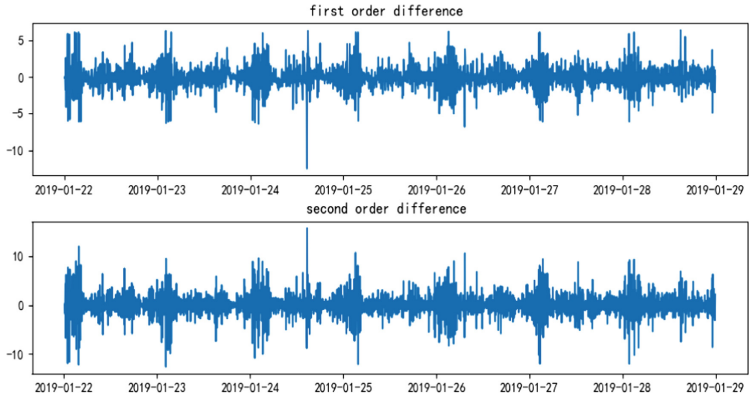

Fig. 1. Differential processing.

As can be seen from Fig. 1, a good stationary can be obtained by applying the second order difference. Therefore, $d = 2$ is set in the model.

Step2: The historical time data of variables are used to predict themselves. However, the premise must satisfy the requirement that the autoregressive model must be of full considerable stationary. AR needs to determine an order p, which means to predict the current value with the historical value of more time in the past. The formula of p autoregression model is defined as Eq. (1).

$$V_{\mathrm{t}} = \mu + \sum_{i=1}^{p} \gamma_i \, V_{t-i} + \varepsilon_t \tag{1}$$

where type of $\mathrm{V_t}$ is the current forecast, μ is constant, p is the order number, γ_i is auto correlation coefficient, ϵ_t is error.

Step3: In order to solve the above AR median error problem, MA is the accumulation of error terms in the autoregression model. The formula definition of q autoregression process is shown in Eq. (2).

$$V_t = \mu + \sum_{i=1}^{q} \varphi_i \varepsilon_{t-i} + \varepsilon_t \tag{2}$$

where type of V_t is the current forecast, μ is constant, q is the order number, φ_i is autocorrelation coefficient, ϵ_t is error.

3 PageRank Traffic Flow Modeling

PageRank algorithm [14] is a method used by Google to measure the rank and importance of web pages. The core idea of this method is that the importance of a web page depends on the quality and quantity of other pages pointing to it. At the beginning, assign an initial PageRank value PR to all web pages, which satisfy Eq. (3).

$$\sum_{i=1}^{p} PR = 1 \tag{3}$$

where the formula (3), p is the total number of pages, and then the PageRank value of each page is iteratively and evenly assigned to the page it points to. In order to solve the situation that the pointing page is 0, the page always assigns the PageRank value to itself, sets a scaling factor to reduce the PageRank value of each node, and the network PageRank value is integrated as 1. In other words, the iterative calculation is carried out through the following Eq. (4).

$$PR_{p_k} = \frac{1-\beta}{n} + \beta \sum_{M_{p_k}} \frac{PR_{p_k}}{|L_{p_j}|} \tag{4}$$

where, $p_1, p_2, \ldots, p_n$ represents web page; M_{p_k} represents the number of pages p_k linked to other pages; L_{p_j} represents the number of links p_j points to other pages. n represents all pages in the network; PR_{p_k} represents the ranking value of page p_k; Beta represents the probability of the user continuing to browse the page, and the beta usually has a probability value of 0.85.

Therefore, PageRank algorithm can solve the problem of data in the form of graph. A random walk model is established on the directed graph, and then the PageRank value of each node in the directed graph and the contribution value (weight) between nodes are calculated iteratively. Based on this theory, this paper proposes the application of PageRank algorithm in solving the prediction of road traffic. Each section of the road network can be set as a node (the node in the directed graph), and the section and the edge constructed by the section can be set as a path (a sequence passed from one node to another node in the directed graph is called a path), and the sum of the paths is the length. In this way, we obtained the directed graph composed of the road network and the required information (end point, path and length), and then used PageRank for modeling. This paper used PageRank to build the traffic model into the following steps.

3.1 Construction of Directed Graph of Road Network

In this paper, the actual road condition information is modeled and processed, and the road condition information diagram is shown in Fig. 2.

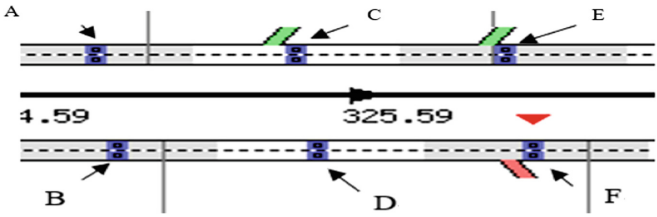

Fig. 2. Road traffic.

As is shown in Fig. 2, road conditions are composed of upstream and downstream. Upstream is composed of section A, section C and section E and the downstream is composed of section B, section D and section F. section A and section F belong to predicted sections. In order to improve the prediction, at the same time, the upstream and downstream traffic flow is predicted in this experiment. The upstream prediction point is A, and the relevant sections are C and E. The downstream prediction point is F, and the relevant sections are B and D. According to the relationship between upstream and downstream sections, the PageRank model is used to construct the upstream directed graph and downstream directed graph, as is shown in Fig. 3 and Fig. 4.

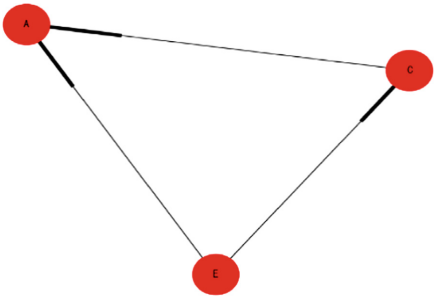

Fig. 3. Upstream directed graph

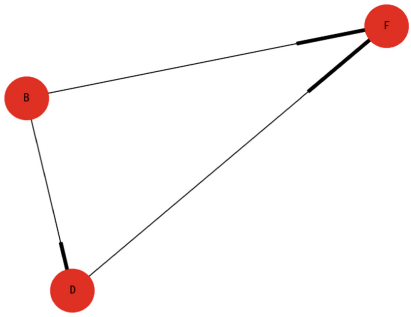

Fig. 4. Downstream directed graph

Figure 3 and Fig. 4 represent the upstream and downstream directed graphs respectively.

The above diagram is explained as follows:

(1) Since upstream *E* will pass through *C* and *A*, when predicting *A*, it is necessary to consider the influence of point *E* on its traffic flow at the next moment, influence of point *E* on its traffic flow at the next moment.
(2) This paper predicts sections *A* and *F*, so there is no degree of points *A* and *F*. In order to display the road network information more intuitively, the section name is

used to represent the directed graph of the road network by the adjacency matrix, and the matrix of the directed graph of the road network is set as G = (V, E), which has vertices and edges. The adjacency matrix of G has the following properties.

$$A(i\ ,\ j) = \begin{cases} 1\ \langle Vi, Vj \rangle \\ 0\ \langle Vi, Vj \rangle \end{cases} \tag{5}$$

The upstream adjacency matrix ($V1, V2, V3$ represents E, C, A) is shown in equation

$$\begin{array}{c|ccc} V1 & 0 & 1 & 1 \\ V2 & 0 & 0 & 1 \\ V3 & 0 & 0 & 0 \end{array} \tag{6}$$

The upstream adjacency matrix ($V1, V2, V3$ represents E, C, A) is shown in equation.

$$\begin{array}{c|ccc} V1 & 0 & 1 & 1 \\ V2 & 0 & 0 & 1 \\ V3 & 0 & 0 & 0 \end{array} \tag{7}$$

3.2 Build a Random Walk Model on the Directed Graph of Road Network

In directed graph defined on the random walk model representation in the directed graph node to another node state transition, however, and all through the formation of the state transition directed edge constitute an order matrix M, so it calculate the transition probability of between two nodes, the network link between the contribution values of calculation, we will know every specific sections of diversion to other sections of situation by contribution value. Specifically, this paper aims to solve shunting, so the specific steps of solving shunting transfer matrix are shown in Eqs. (8), (9) and (10).

$$F = \left[f_{ij}\right]_{n*n} \tag{8}$$

$$f_{ij} \geq 0 \tag{9}$$

$$\sum_{i=1}^{n} f_{ij} = 1 \tag{10}$$

where f_{ij} refers to the node j points to the node i; otherwise, it is 0, I, $j = 1\ldots$ N. According to the establishment of the above directed graph, we can get the upstream and downstream shunt transfer matrix, which is shown in Eq. (11).

$$F1 = \begin{array}{|ccc|} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{array} \tag{11}$$

Similarly, the downstream shunt transfer matrix is shown in Eq. (12).

$$F1=\begin{vmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{vmatrix} \tag{12}$$

In the table, the vertical axis represents the probability value of branching from other sections, and the horizontal coordinate represents the probability value of between every two sections, which all represent the probability distribution value of the state transition matrix. With this probability value, the distributary ratio between sections is solved. This is a key step in this article.

3.3 The Stationary Probability Distribution Value Is Solved by Power Method

In order to make the probability distribution tend to be stable and ensure the accuracy of the shunt ratio, the solution needs to be solved by power method. The state transition matrix can be expressed as shown in Eq. (13).

$$R = (dM + \frac{1-d}{n}E)R \tag{13}$$

where d is the damped silver and E is the unit vector.

According to the directed graph and diversion and transfer evidence, the PR value in the directed graph of road network can be calculated.

We're going to end up with a stationary probability distribution. In addition, we use formula (14) to calculate the traffic ranking value of each road section (small value means no cars are usually on this road section), and we can choose the travel route according to this traffic ranking value.

$$PR(v_i) \geq 0, i = 1 \ldots 6 \tag{14}$$

$$\sum_{i=1}^{n} PR(v_i) = 1 \tag{15}$$

$$PR(v_i) = \sum_{Vj \in M_{(v_j)}}^{n} \frac{PR(v_i)}{L_{(v_j)}} \tag{16}$$

where $M(v_i)$ represents the node set of v_j, and $L(v_j)$ represents the number of directed edges of node v_j pointing to other road segments.

4 Experiments

4.1 Data Collection

The experimental data in this paper are from January 22 to January 29 in PEMS database. Due to the large number of sections on the road section, six points were randomly selected in the road section for better performance, and the six points were numbered as {a-f} respectively. The speed and traffic data of the six sections are shown in Fig. 5 and Fig. 6.

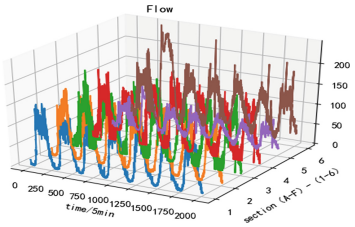

Fig. 5. Speed

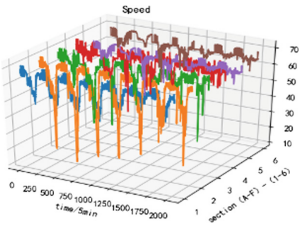

Fig. 6. Flow

As can be seen in the Fig. 5 and Fig. 6, the data shows periodic changes, which is conducive to the ARIMA model. The following Table 1 shows some of the input data.

Table 1. Input data

5 min	Speed	Flow
1/22/2019 10:35	66.60	99
1/22/2019 10:40	66.70	103
1/22/2019 10:45	67.30	100
1/22/2019 10:50	66.90	124
1/22/2019 10:55	66.30	91
1/22/2019 11:00	64.80	102
1/22/2019 11:05	65.90	122
1/22/2019 11:10	65.20	113
1/22/2019 11:15	65.20	119
1/22/2019 11:20	65.60	95
1/22/2019 11:25	66.50	105
1/22/2019 11:30	65.00	92
1/22/2019 11:35	65.50	98
1/22/2019 11:40	64.60	92
1/22/2019 11:45	66.10	90

As can be seen from Table 1, the data collected on January 22. 5 min is the granularity of data collection; Speed is average Speed; Flow is average traffic Flow.

4.2 ARIMA Model Predicts Results

The Arima prediction model was obtained through the training of historical data, and the predicted values (speed and flow) were used as the input values of the PageRank model. The predicted road sections in this paper are A and F, so the predicted traffic data of each road section in the road network should be obtained before the experiment. Then, according to the PageRank model, the weights between sections were obtained to get the final prediction results.

Traffic data (speed and flow) from January 22 to January 28 were used as the training set and data of Jan. 29 as the test set. The speed and flow prediction results of ARIMA model for the six sections are shown in Fig. 7 and Fig. 8.

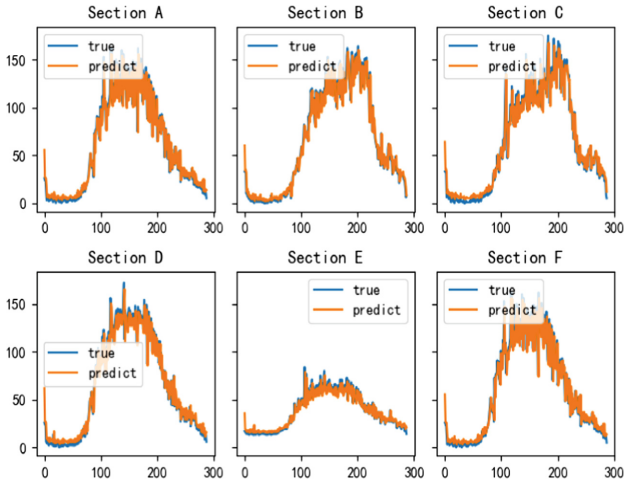

Fig. 7. Flow prediction result

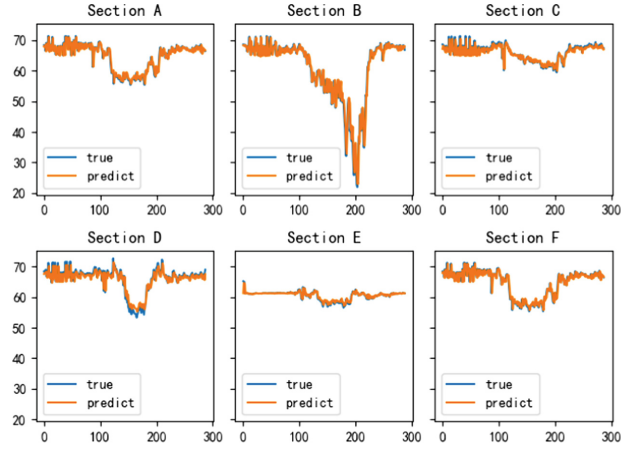

Fig. 8. Speed prediction result.

It can be seen from Fig. 7 and Fig. 8 that the predicted effect is consistent with the actual value.

4.2.1 Evaluation Index of ARIMA Model

In order to prove the accuracy of prediction ability of ARIMA model from the perspective of quantification, this paper introduces three evaluation indexes: average absolute percentage error (MAPE), mean square error (MSE) and R^2 (the closer to 1, the better) to verify the quality of ARIMA model.

The average absolute percentage error, mean square error and regression coefficient R^2 of the three evaluation indexes were calculated in the experiment, and the results are shown in Table 2.

Table 2. The evaluation indexes

Index	MSE(Speed/flow)	MAPE (Speed/flow)	R^2 (Speed/flow)
Results	0.21/0.19	0.22/0.17	0.9/0.98

In summary, it can be seen from the results of the evaluation indexes that the results are relatively good regardless of MSE, MAPE or R^2. Therefore, it can be considered that ARIMA model has a good predictive ability for the experiment in this paper.

MSE residuals of flow and speed are shown in Fig. 9 and Fig. 10.

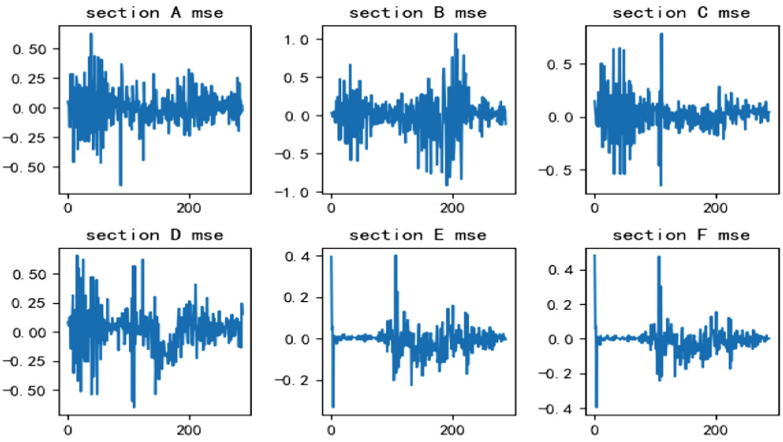

Fig. 9. Flow residual

As is shown in Fig. 9 and Fig. 10, the values of the residuals are relatively small, so it can be seen that the real value and the predicted value are very close, thus verifying the reliability of the model.

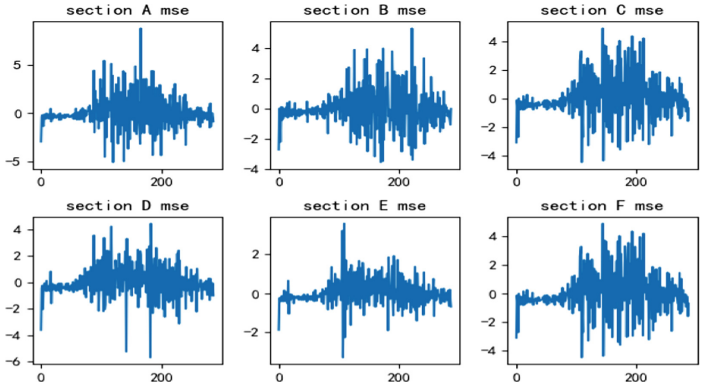

Fig. 10. Speed residual

4.3 PageRank Traffic Flow Prediction Results

4.3.1 Traffic Forecast

The predicted values of segment A and f can be calculated through the predicted values calculated by Arima model and the weights calculated by PageRank model, and the predicted results are shown in Fig. 11.

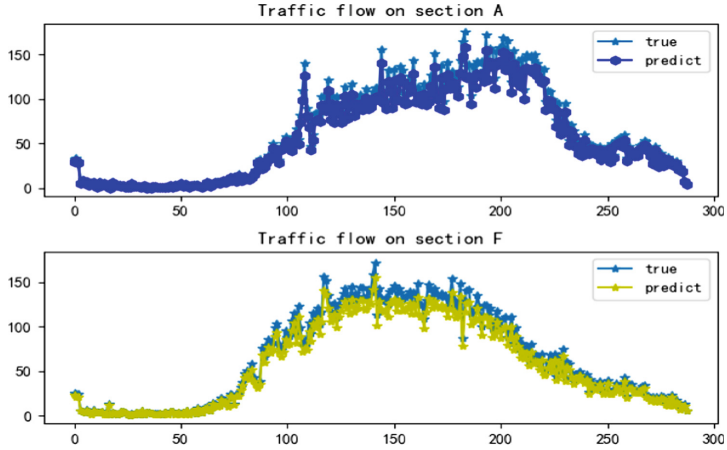

Fig. 11. The result of prediction.

As can be seen from Fig. 11, the combined model achieves a good effect in prediction. Therefore, this combined model has realized the prediction of the road network by utilizing the degree of correlation between sections.

4.3.2 Time Forecast

The velocity V is calculated by ARIMA model, so we can calculate the time T of each road segment to the predicted point by the distance L. The predicted time is shown in Fig. 12.

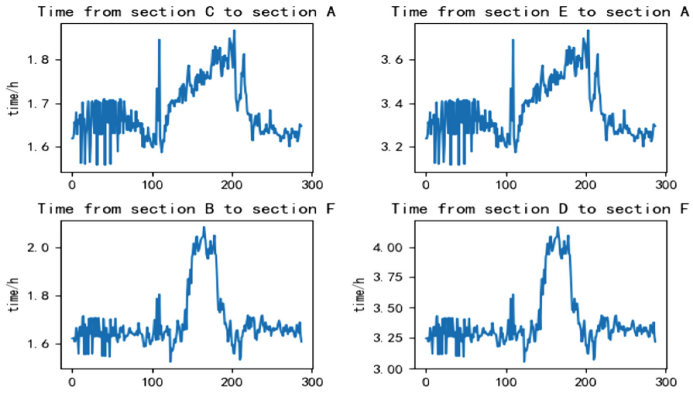

Fig. 12. Time prediction.

$$T = \frac{L}{V} \tag{17}$$

where (17) L represents the length of each section of expressway, T is the time between each sections.

When a certain road is predicted to be congested, road workers can work timely according to the time so as to ease the degree of congestion.

5 Discussions

Different from the existing traffic flow prediction models, a combination method based on Arima model and PageRank model is proposed to promote the whole network forecast. Compared with the traditional prediction methods, the combined model method proposed in this paper improves the prediction accuracy and application. Its value is mainly reflected in the following aspects:

The prediction accuracy reaches 94.7%, which is better than other combined models. Many methods are aimed at single point prediction, but single point prediction has some limitations. Therefore, this method fundamentally breaks through the bottleneck of single point prediction and realizes the prediction of the whole road network. The PageRank model can simulate the road network information map well, for example, the degree of linkage between parts.

6 Conclusions and Future Work

This method not only solves the prediction problem of the whole road network, but also serves as a warning to the road network planning. Therefore, it plays a certain role in solving traffic flow prediction. However, in order to better solve the complex road network diagram, more factors need to be considered, such as the performance of the algorithm needs to be improved, the road network information is more complex and so on. The next step of this paper is to solve intelligent transportation problems based on big data platform.

Acknowledgement. The work was supported by the Graduate Innovation and Entrepreneurship Program in Shanghai University in China under Grant No. 2019GY04.

References

1. Cheol, O., Jun-seok, O., Ritchie, S.G.: Real-time hazardous traffic condition warning system: framework and evaluation. IEEE Trans. Intell. Transp. Syst. **6**(3), 265–272 (2005)
2. Ding, W., Gong, Y., Nan, N.: Toward cognitive vehicles. IEEE Intell. Syst. **26**(3), 76–80 (2011)
3. Wu, C., Zhao, G., Ou, B.: A fuel economy optimization system with applications in vehicles with human drivers and autonomous vehicles. Transp. Res. Part D Transp. Environ. **16**(7), 515–524 (2011)
4. Luettel, T., Himmelsbach, M., Wuensche, H.J.: Autonomous ground vehicles-concepts and a path to the future. Proc. IEEE **100**, 1831–1839 (2012)
5. Grant-muller, S., Usher, M.: Intelligent transport systems: the propensity for environmental and economic benefits. Technol. Forecast. Soc. Chang. **8**(2), 149–166 (2014)
6. Nair, A.S., Liu, J-C., Rileft, L, et al.: Non-linear analysis of traffic flow. In: Proceedings of the Intelligent Transportation Systems (2001)
7. Hua, J., Faghri, A.: Dynamic traffic pattern classification using artificial neural networks. The TRIS and ITRD Database (1993)
8. Wei, W., Wu, H., Ma, H.: An Auto Encoder and LSTM-Based Traffic Flow Prediction Method. Sensors (Basel, Switzerland) **19**(13), 2946 (2019)
9. Davis, G.A., Nihan, N.L.: Nonparametric regression and short-term freeway traffic forecasting. J. Transp. Eng. **117**(2), 178–188 (1991)
10. Dougherty, M.S., Cobbett, M.R.: Short-term inter-urban traffic forecasts using neural networks. Int. J. Forecast. **13**(1), 21–23 (1997)
11. Abdulhal, B., Ritchie, S.G.: Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. Transp. Res. Part C Emerg. Technol. **7**(5), 26–280 (1999)
12. Park, B.B.: Hybrid neuro-fuzzy application in short-term freeway traffic volume forecasting. Transp. Res. Rec. J. Transp. Res. Board **1802**(1), 190–196 (2002)
13. Tang, J., Chen, X., Hu, Z., Zong, F., Han, C., Li, L.: Traffic flow prediction based on combination of support vector machine and data denoising schemes. Physica A Stat. Mech. Appl. **534**, 120642 (2019)
14. Payandeh, S., Chiu, E.: Application of modified pagerank algorithm for anomaly detection in movements of older adults. Int. J. Telemed. Appl. **2019**, 1–9 (2019)