

Root Cause Identification for Road Network Congestion Using the Gradient Boosting Decision Trees

Yue Chen*, Changle Li*[†], Senior Member, IEEE, Wenwei Yue*, Student Member, IEEE, Hehe Zhang* and Guoqiang Mao*, Fellow, IEEE

*State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, Shaanxi 710071, China

[†]Email: clli@mail.xidian.edu.cn

Abstract—Identifying the root cause in urban road networks and ranking the influential factors can benefit traffic management for improving traffic condition. Traditional congestion identification studies paid attention to identify traffic bottlenecks, namely the most vulnerable points in a road network, without consideration of root causes that leading to the congestion. In this paper, we propose a gradient boosting decision trees (GBDTs) based method to identify the root cause of road network congestion and rank the influential factors using different types of explanatory variables. Based on Sioux Falls network, different signal control strategies at intersections and number of lanes on road segments under different traffic flows are conducted as samples using Simulation of Urban Mobility (SUMO) to train and test the GBDT model. Simulation results indicate that the GBDT model can achieve superior performance in average travel speed prediction and identify the root causes of congestion by prioritizing the relative importance of influential factors, such as lane numbers and signal control strategies, compared with other algorithms.

Index Terms—Root cause of congestion, GBDT, SUMO

I. INTRODUCTION

Traffic congestion is an increasingly serious problem that nearly all cities face, especially in metropolis. First, traffic congestion causes a rise of trip time cost, which takes economical losses due to reduction of production time. Second, traffic congestion causes a rise of fuel consumption and then causes the rise of carbon and oxynitride emissions which aggravates atmospheric pollution and greenhouse effect. Last but not the least, with the growing increase of the retention of vehicles, congestion is becoming increasingly fierce [1], [2], [3].

As the major congestion contributors in a road network, traffic bottleneck identification has attracted much attention recently [4]. Li *et al.* [5] developed a method with the combination of graph theory and Markov analysis to identify urban bottlenecks. In [6], Ma *et al.* defined a parameters I_m based on traffic impedance C_{rs} and network effectiveness E . They compared the parameter I_m before and after a particular road segment failure (in congestion) and regarded the road segment with more difference of parameter I_m as a bottleneck. Ye *et al.* [7] used a critical index v/c based on the ratio of traffic flow and road capacity of a road segment to identify whether a road segment is a bottleneck or not. Lee *et al.* [8] developed a three-phase spatio-temporal bottleneck mining model to

identify bottlenecks in urban road networks and considered that bottlenecks most likely existed in the spatial cross section of two congestion propagation patterns. In summary, these works can identify the most congestion contributed points in a road network, however, the root causes that leading to congestion are neglected. If the root causes of congestion can be identified effectively, more efficient strategies can be applied to relieve congestion in the entire road network.

To fill the gap, in this paper, we utilize GBDT based approach to identify the root cause of congestion in road networks. Firstly, we propose a gradient boosting decision trees (GBDTs) based method to model and predict the performance of road networks with the target of average travel speed from explanatory variables corresponding to the factors that can be collected easily (*e.g.*, lane numbers, signal control strategy and traffic flows). Secondly, based on the GBDT model, the influential factors that influence the performance of road networks are prioritized and the major factors can be given to identify the root cause of congestion. By choosing the major factors as the root cause of congestion, strategies can be implemented to relieve congestion more efficiently. Finally, a simulation based on SUMO is used to illustrate the effectiveness of proposed root cause of congestion identification method. More specifically, contributions of this paper are presented as follows:

- A gradient boosting decision trees based method is used to model and predict the average travel speed of road networks according to easily measuring influential factors such as lane numbers, signal control strategies and traffic flows, which can better quantify the different influence of these factors on road network fluency and further estimate the traffic condition of entire road networks.
- A prioritization of the influential factors is utilized to identify the root cause of congestion which demonstrates the influence of each factor on the average travel speed of road networks and provides a reliable method to locate the root causes of congestion.
- Simulations are conducted using Simulation of Urban Mobility (SUMO) where different signal control strategies at intersections and number of lanes on road seg-

ments under different traffic flows are encoded as samples to train and test the GBDT model. A comparison between GBDT method and other algorithms are also developed, which validates the effectiveness of GBDT method compared with other algorithms.

The rest of paper is organized as follows. In Section II, we utilize GBDT model to predict the average travel speed of road networks and prioritize the relative importance for congestion influential factors. Based on a traffic simulator SUMO. Section III presents the data resources used in this study. The simulation results of the proposed model and a comparison with other algorithms are conducted in section IV. Conclusion is outlined at the end.

II. METHODOLOGY

A. Gradient Boosting Decision Trees

In this study, a type of machine learning method called gradient boosting decision trees (GBDT) is used to model and predict the average travel speed of a road network. Assuming that $F(x)$ is an approximation of the label y based on a set of predictor variables x , the least square error function is applied as the loss function to estimate the approximation function as follow [9], [10]:

$$L(y, F) = \frac{1}{2}[y - F]^2. \quad (1)$$

Assuming that the number of splits is J for each sub-tree, which splits the input space into J regions just like $R_{1m}, R_{2m}, \dots, R_{jm}$ and predicts a constant value b_{jm} into region R_{jm} . Thus, each decision tree can be written as follow [11], [17]:

$$h_m(x) = \sum_{j=1}^J b_{jm}I, \quad (2)$$

where $I = 1$ if $x \in R_{jm}$; $I = 0$ otherwise. Considering the data: $\{y_i, x_i\}_1^N$, the gradient boosting decision tree iteratively generates M different regression trees $h_1(x), \dots, h_M(x)$. The updating form function $F_m(x)$ is given with a gradient descent step size ρ_m as follows [12], [13], [14]:

$$F_m(x) = F_{m-1}(x) + \rho_m \sum_{j=1}^J b_{jm}I(x \in R_{jm}). \quad (3)$$

$$\rho_m = \arg \min_{\rho} \sum_{i=1}^N L(y_i, F_{m-1}(x_i) + \rho \sum_{j=1}^J b_{jm}I(x \in R_{jm})). \quad (4)$$

Finding an optimal partition γ_{jm} for each region R_{jm} , then the (3) can be presented without b_{jm} as follows [9], [12]:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^J \gamma_{jm}I(x \in R_{jm}), \quad (5)$$

and to obtain the optimal can be on the basis as follows:

$$\begin{aligned} \gamma_m &= \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \\ &= \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} (\tilde{y} - \gamma)^2, \end{aligned} \quad (6)$$

where

$$\tilde{y}_i = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F_m(x) = F_{m-1}(x)}. \quad (7)$$

The gradient boosting trees build the model step by step and update the parameter by minimizing the value of certain loss function. To prevent over-fitting and improve the model performance, it applies a strategy to scale the contribution of base tree with a learning rate ξ ($0 < \xi < 1$) [9], [15]. Thus, (5) can be written as below:

$$F_m(x) = F_{m-1}(x) + \xi \sum_{j=1}^J \gamma_{jm}I(x \in R_{jm}). \quad (8)$$

Choosing a small learning rate can better minimize the loss function but may add a large number of trees to the model. Thus, the complexity of each sub-tree should be limited to obtain a good cost-effectiveness model at the balance between complex interactions capture and model complexity. Selecting the combination of parameters, the GBDT model with an optimal performance can be found.

B. Relative Importance of Influential Factors

The GBDT method can identify and rank the influences of predictor variables on response predictions. For a single decision tree T , relative importance of the predictor x_k has an approximation in predicting the response as follow [16]:

$$I_K^2(T) = \sum_{t=1}^{J-1} \tilde{\tau}_t^2 I(v(t) = k), \quad (9)$$

where the summation over the non-terminal nodes t of J -terminal node tree T , x_k is the splitting variable associated with node t , and $\tilde{\tau}_t^2$ is the corresponding empirical improvement in the form of squared error as a result of using variable x_k as a splitting variable at the non-terminal node t . For a collection of decision trees $\{T_m\}_1^M$, (9) can be represented by its average over all of the sub-trees[17]:

$$I_K^2(T) = \frac{1}{M} \sum_{m=1}^M I_K^2(T_m). \quad (10)$$

III. DATA SOURCES

The data used in this study are from simulation based on SUMO. It was simulated on a simplified network based on the City of Sioux Falls, South Dakota, USA (the road network has 76 edges and 24 intersections as shown in Figure 2). Each sample is a vector including traffic flow and lane number of each edge and signal control strategy of each intersection corresponding to a random initialization of road network elements (lane numbers and signal control strategy).

TABLE I
DESCRIPTION OF INDEPENDENT VARIABLES USED IN GBDT

| Variables | Value set |
|--------------|--------------------------|
| Lane number | 1 = single lane edge |
| | 2 = double lane edge |
| | 3 = three-lane edge |
| Intersection | 0.1 = without control |
| | 0.5 = fixed-time control |
| | 1 = responsive control |
| Traffic flow | R^+ |

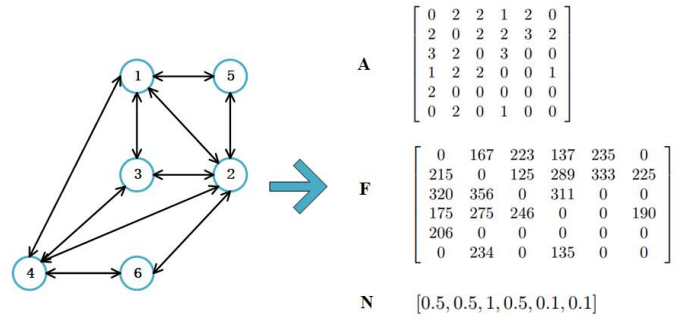


Fig. 1. Example of getting samples from a road network.

Edges in the road network can have 1 to 3 lanes and traffic lights have 3 types: without control, fixed-time control and actuated type. A matrix A is used to describe the road network after a lane numbers initialization: A_{ij} is the lane number in the link from node i to node j (if there is no link, the value should be 0) and A_{ii} is set to be 0. A vector \vec{N} is used to describe the signal control strategy of intersections. \vec{N}_i represents the signal control strategy of node i (without control: 0.1, fixed-time control: 0.5, responsive control: 1). A matrix F is used to describe the traffic flow on edges: $F_{i,j}$ represents the traffic flow during the simulation. The traffic flow was collected and converted to hourly volume based on a interface called Traci which links and controls the SUMO server. Samples are obtained by flattening the matrix A , F to vectors with deleting the zeros and splicing it with vector \vec{N} . It has the form as follow:

$$\left\{ \vec{A}', \vec{F}', \vec{N} \right\}^T, \quad (11)$$

where \vec{A}' and \vec{F}' represent the vectors converted from matrix A and B . Label data is the average travel speed during each simulation in road network.

Each sample of the data set used in this study has 176 dimensions (traffic flow in 76 edges, lane numbers of 76 edges and signal control strategies of 24 intersections). An example of getting samples is shown in Figure 1. The label data is the mean value of all simulation step's average travel speed of road network which represents the congestion level in road network which can be represented as follow:

$$l = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{v_j}{m_i}, \quad (12)$$

where n is the total simulation steps and m_i is the vehicle number in the road network at the step i . v_j is the speed of vehicle j .

IV. SIMULATION RESULT

A. Optimization of Model Parameter

For the performance of predicting the average travel speed in the road network, the determination coefficient (R^2 score) is used as the evaluation criterion in this study. The determination coefficient is defined as follows:

$$v = \sum_i^m (\hat{y}_i - \tilde{y})^2, \quad (13)$$

$$u = \sum_i^m (\hat{y}_i - y_i)^2, \quad (14)$$

$$R^2(y, \hat{y}) = 1 - \frac{u}{v}, \quad (15)$$

where \hat{y}_i is the i^{th} data in label dataset, y_i is the corresponding predicted value and \tilde{y} is the mean value of the label data. In addition, m is the sample numbers. In this study, 80% of sample data are used for training while 20% of sample data are used for testing. Generally, a value of R^2 score close to 1 indicates a high performance of the model and the value can be negative because the model can be arbitrarily worse. To test the model performance of different combinations of parameters, several GBDT models are built with various learning rates (ξ values from 0.01 to 1 with a step of 0.01), tree max depth (d values from 1 to 50 with a step of 1) and estimator number (values from 1 to 200). To determine the optimal parameters, the estimator numbers are tested firstly. Estimator numbers represent the maximal number of weak learning regressor. Generally, a large estimator number will make the model over-fitting while a small estimator number will lead to the model under-fitting. As is shown in Figure 3, it indicates that the score tends to be stable at the value of 1.0 on training set and stable nearby the value of 0.8 on test set after 50 iterations with a fixed learning rate of 0.01. The score is stable at the value of 1 on training set after 400 iterations and stable at the value of 0.8 on test set after 600 iterations with a fixed learning rate 0.1. The last test with a fixed learning rate 0.5 indicates that the score is stable at the value of 1.0 only after 10 iterations on training set and shocks under the value of 0.8. For the parameter of learning rate, that is the weight reduction coefficient of each weak learning regressor, a smaller value means that it needs more weak learning regressor estimator numbers to obtain the same training effect on training set. With reference to result of above test, in this paper, the learning rate is tested from 0.01 to 1 with estimator numbers of fixed 1000. As the result shown in Figure 4, the model performance reaches its best with a learning rate of 0.05. Considering these

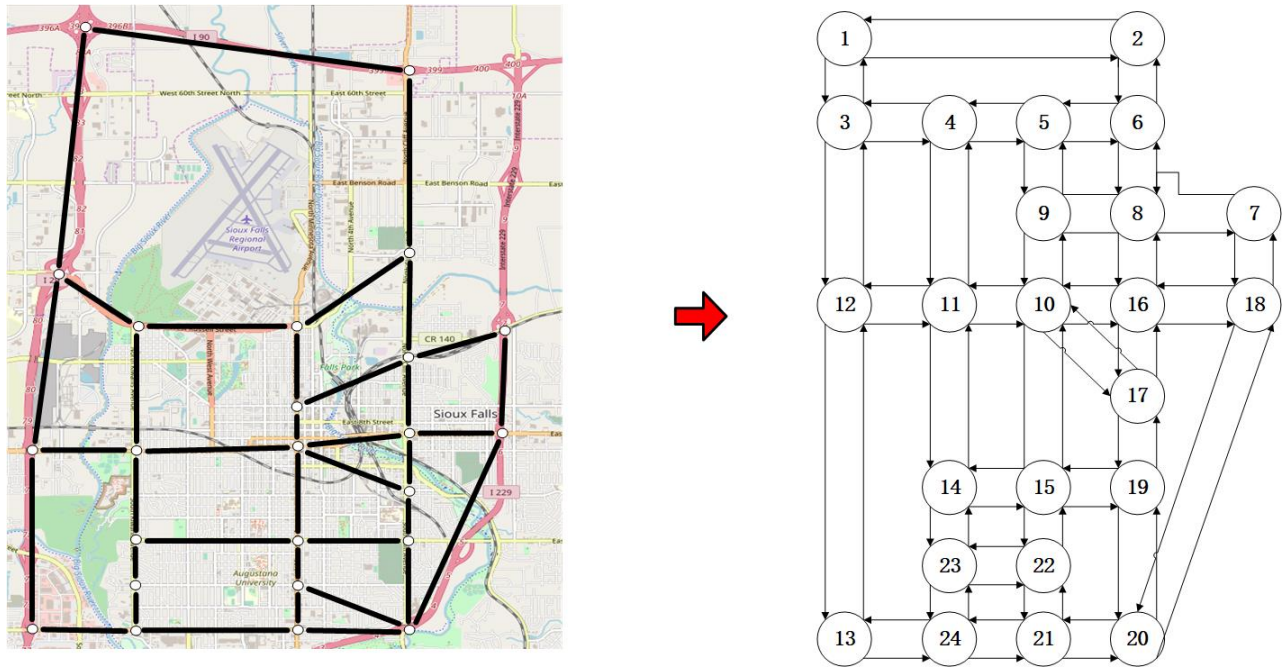


Fig. 2. Sioux Falls network.

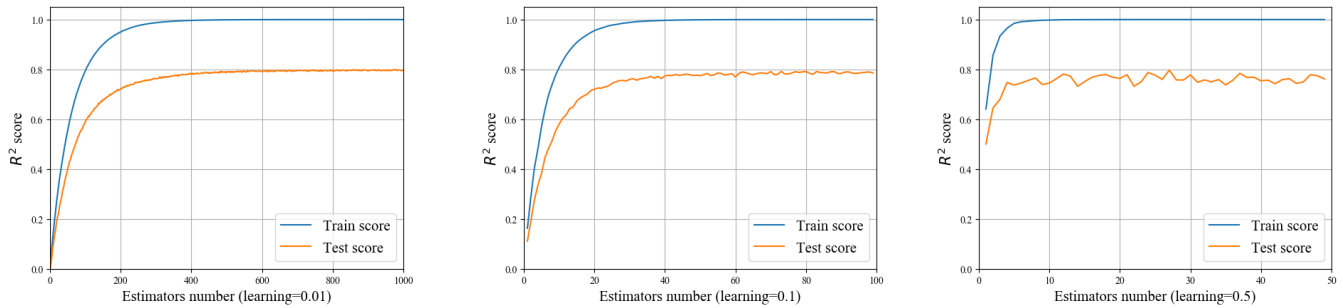


Fig. 3. Influence of estimator number on model performance.

two parameters as a whole, the learning rate was set to be 0.4 and the estimator number was set to be 100.

The parameter of the maximum depth and maximum leaf nodes limit the complexity of sub-trees in GBDT model. Without the limit of complexity of sub-trees, the model will take a large memory space of computer and be worse on generalizations ability, that is the model will be over-fitting. Through experiment shown in Figure 5, the max depth was determined and set to be 3 with the highest R^2 score 0.81 and the max leaf nodes was set to be 8 with the highest R^2 score 0.825 in experiment. By analyzing the result in Figure 3 to Figure 6, the best model performance can be acquired. The optimal model is obtained when learning rate has a value of 0.4, estimator number has a value of 100, maximum leaf nodes has a value of 8 and tree maximum depth values 3. The R^2 score of optimal model on test dataset is 0.864.

B. Comparison with other algorithms

A comparison with other algorithms including multilayer perceptron (MLP) and random forest (RF) is used to demonstrate the effectiveness of GBDT model. It was tested on the subsets of test dataset and each subset is corresponding to a traffic flow (hourly volume). There are total three subsets and corresponding three values of traffic flow: 3600 vehicles per hour, 5400 vehicles per hour and 7200 vehicles per hour. Table 2 shows the comparison result. The multilayer perceptron is a type of back propagation neural network. The multilayer perceptron used in this paper has two hidden layers (the first hidden layer's dimension is 100 and the second is 11) and uses the ReLU activation after each hidden layer. Stochastic gradient descent (SGD) was used to train the network. For RF algorithm, the estimator number and max depth of sub-trees are important parameters the same as GBDT. They are set to be 200 and 3 after search in parameter space.

Comparing the results of three algorithms, it can be seen that

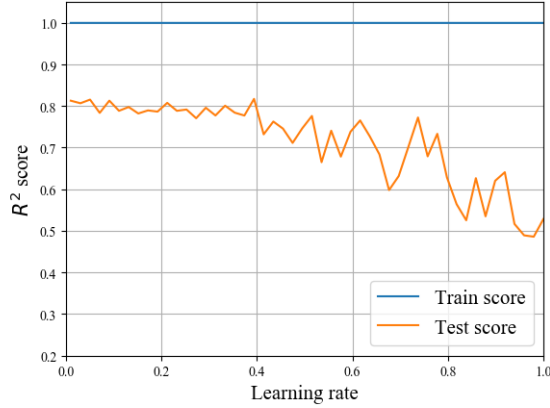


Fig. 4. Influence of learning rate on model performance.

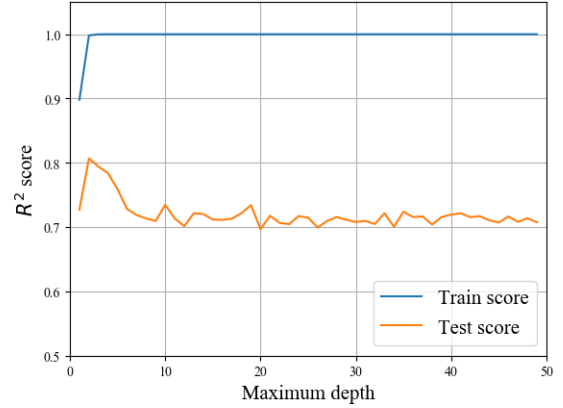


Fig. 6. Influence of max depth on model performance.

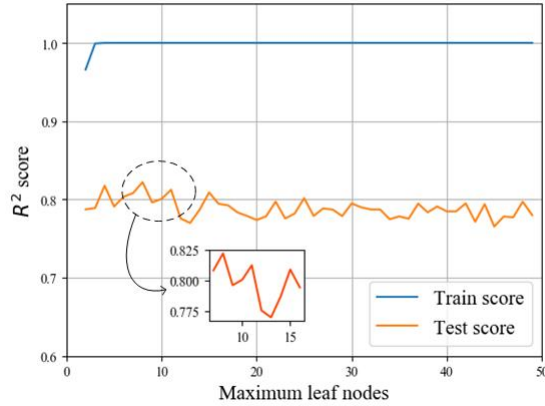


Fig. 5. Influence of maximum leaf nodes on model performance.

GBDT model has the best performance on entire test dataset with a R^2 score of 0.866. On subsets, GBDT model gets the highest score 0.942 when the traffic volume is 3600 vehicles per hour, gets a score of 0.855 when traffic volume is 5400 vehicles per hour and a score of 0.807 when traffic volume is 7200 vehicles per hour. It indicates that GBDT model has a good performance for predicting the average travel speed at different traffic volumes. For RF and MLP, the performance is poor when traffic volume is 7200 vehicles per hour which makes them cannot fit the high volume situation. Numerically, GBDT is 27.9% higher on performance than RF and 39.2% higher than MLP. This result shows the effectiveness of GBDT.

C. Root Cause Identification and Relative Importance Rank

To explore the different influences of each element in the road network on the average travel speed of road network, the relative importance of predictor variables is calculated using the optimal model. A higher value of relative importance indicates a more obvious influence of the predictor variable on the performance of road network. Thus, the root causes of congestion in a certain road network can be identified ac-

TABLE II
COMPARISON WITH OTHER ALGORITHMS

| Traffic volume of subset | Prediction performance (R^2 score) | | |
|--------------------------|---------------------------------------|-------|-------|
| | RF | MLP | GBDT |
| 3600 | 0.733 | 0.786 | 0.942 |
| 5400 | 0.696 | 0.639 | 0.855 |
| 7200 | 0.358 | 0.455 | 0.807 |
| Entire test dataset | 0.677 | 0.622 | 0.866 |

ording to their relative importance rank. The result of relative importance from our optimal GBDT model is shown in Figure 7. It can be seen that the most influential factor (the traffic flow of edge A) has a relative importance of 23.3%, the most influential intersection has a relative importance of 4.9%. As shown in Table 3, in general, traffic flow is the most important contributor to the average travel speed of road network with a total relative importance of 80.2%, signal control strategy of intersection contributes a total relative importance of 15.3% and lane number of road segment contributes a total relative importance of 4.7%.

To identify the root cause of congestion in road networks, the influential factors with the highest three values of relative importance are considered as the root cause of congestion. As shown in Figure 8, improving traffic control strategy at node 10 from fixed-time control to responsive control and increasing the road capacity on edge A and B can increase the average travel speed by 36.2% in average of entire road networks.

V. CONCLUSION

Reliable prediction of average travel speed of road networks is of vital importance for identifying the root cause of congestion. This study contributes to model and predicts the average travel speed of road networks using a gradient boosting decision tree (GBDT) based method. The GBDT based method has a perfect performance on the test set and rank the relative importance of each variable automatically which benefits to identify the root cause of congestion. Data including traffic flow, lane number and signal control strategy on each edge and node from simulation based on SUMO

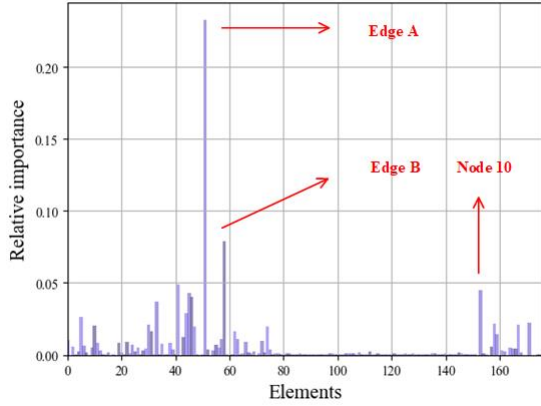


Fig. 7. Relative importance of elements in road network.

TABLE III
RELATIVE IMPORTANCE OF DIFFERENT ELEMENTS IN ROAD NETWORK

| Categories | Variables | Rank | Relative importance |
|--------------|-------------------------|------|---------------------|
| Traffic flow | Hourly traffic volume | 1 | 80.0% |
| Intersection | signal control strategy | 2 | 15.3% |
| Road segment | Lane number | 3 | 4.7% |

are used to verify the effectiveness of GBDT algorithm. Comparison results show that the GBDT model has a better performance than other algorithms. From the optimal GBDT model, the relative importance of each variable is calculated to identify the root cause of congestion.

The developed method in this paper can benefit to generate more efficient strategies to relieve the congestion in urban road networks according to the root cause identified by our GBDT model. In the future, more detailed traffic information such as road segment lengths, betweenness and centrality will be included to analyze their influence on traffic condition and identify root causes of congestion in road networks.

VI. ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (2019YFB1600100), National Natural Science Foundation of China (U1801266), Key R&D Program of Shaanxi (2018ZDXM-GY-038, 2018ZDCXL-GY-04-02), the Youth Innovation Team of Shaanxi Universities, and the Science and Technology Projects of Xi'an, China (201809170CX11JC12).

REFERENCES

- [1] L. Morgan, "The effects of traffic congestion," Sep. 2014. [Online]. Available: <http://traveltips.usatoday.com/effects-traffic-congestion-61043.html>
- [2] A. Ahmad, R. Arshad, S. Mahmud, G. Khan, and H. Al-Raweshidy, "Earliest-deadline-based scheduling to reduce urban traffic congestion," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1510–1526, Aug. 2014.
- [3] W. Yue, C. Li, S. Wang, Z. Xu and G. Mao, "Towards enhanced recovery and system stability: Analytical solutions for dynamic incident effects in road networks," *IEEE Trans. Intell. Transp. Syst.*, Early Access, 2020.

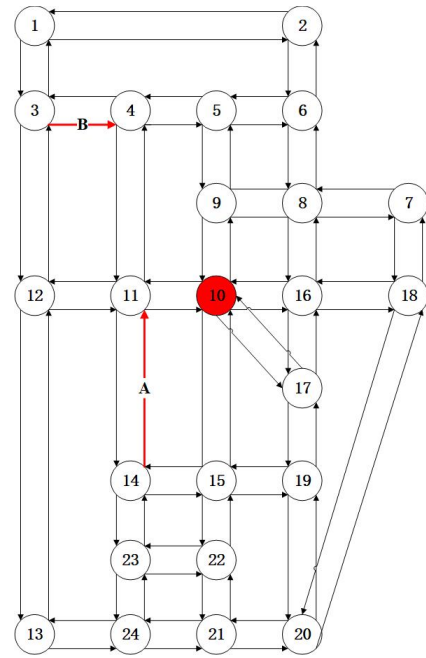


Fig. 8. Root cause of congestion identification.

- [4] D. Hale, R. Jagannathan, M. Xyntarakis, P. Su, X. Jiang, J. Ma, J. Hu, and C. Krause, "Traffic Bottlenecks: Identification and Solutions," Federal Highway Admin. (FHWA), *U.S. Dept. Transp., Washington, D.C., USA, Tech. Rep. FHWA-HRT-16-036*, 2016.
- [5] C. Li, W. Yue, G. Mao, and Z. Xu, "Congestion propagation based bottleneck identification in urban road networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 4827–4841, 2020.
- [6] J. Ma, C. Li, Z. Liu, and Y. Duan, "On traffic bottleneck in green ITS navigation: An identification method," in *Proc. IEEE Veh. Technol. Conf.*, 2016, pp. 1–5.
- [7] X. Ye, S. Deng, W. Yang, Z. Bao, and J. Chen, "Evaluating the impacts of travel information on urban traffic congestion propagation and bottleneck identification," in *Proc. Cota Int. Conf. Transp. Professionals*, 2014, pp. 1890–1901.
- [8] W. Lee, S. Tseng, J. Shieh, and H. Chen, "Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1047–1056, Dec. 2011.
- [9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [10] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.
- [11] G. De'ath, "Boosted trees for ecological modeling and prediction," *Ecol. Soc. Amer.*, vol. 88, no. 1, pp. 243–251, Jan. 2007.
- [12] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015.
- [13] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," 2nd ed. *New York, NY, USA: Springer*, 2009.
- [14] C. Ding, X. Wu, G. Yu, and Y. Wang, "A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data," *Transp. Res. C, Emerg. Technol.*, vol. 72, pp. 225–238, Nov. 2016.
- [15] M. Schonlau, "Boosted regression (boosting): An introductory tutorial and a Stata plugin," *Stata J.*, vol. 5, no. 3, pp. 330–354, 2005.
- [16] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and Regression Trees," *Boca Raton, FL, USA: CRC Press*, Nov. 1984.
- [17] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2303–2310, Sep. 2017.