

Traffic state prediction based on stochastic repetitive hill climbing method

Z. Shen (✉ shenzhou.research@gmail.com)

Research Article

Keywords: Traffic state prediction, hill climbing method

Posted Date: October 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1006289/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Traffic state prediction based on stochastic repetitive hill climbing method

Zhou Shen*

Abstract

A reasonable structure of traffic state network is a prerequisite for traffic state prediction. In order to overcome the shortcomings of the hill climbing method, a traffic state prediction method based on the random repetitive hill climbing method is proposed. A multi-network structure is obtained by iteratively running the hill-climbing method on the randomly generated directed acyclic graph; the node and directed edge selection criteria in the optimal Bayesian network structure are determined by defining the confidence degree of directed edges and calculating the confidence threshold; using the optimal Bayesian network structure, four traffic states, such as smooth, smooth, congested and blocked, are predicted and evaluated comprehensively. The analysis results show that the overall accuracy of the method for traffic state prediction exceeds 85% when only two variables such as time of day and holiday are selected, which can provide effective methods and data support for highway operation state monitoring and early warning and decision analysis.

1 Introduction

The complete acquisition, real-time identification and accurate prediction of highway traffic state is the basis for accurately grasping the behavior of traffic system, making scientific traffic management decisions and giving full play to the potential of traffic facilities on the one hand [1, 2]; on the other hand, it is also an important theoretical basis for analyzing the mechanism of traffic state evolution and grasping the law of traffic congestion occurrence [3].

Bayesian Network (BN) was first introduced by Pearl in 1986 in expert system, which is an ideal model for data mining and uncertainty knowledge representation [4]. In the field of intelligent transportation, the application of Bayesian network for uncertainty inference analysis on travel behavior, accident causation and traffic flow prediction is a hot research topic at home and abroad [5]. The existing research mainly focuses on Bayesian network learning and network inference algorithm implementation, but the learning algorithm for

*The authors declare no competing interests.

Bayesian network structure is relatively single, and the research on the reasonable trade-offs of network nodes and directed edges is not deep enough to verify the advantages and disadvantages of the network structure [6]. In this paper, we propose a traffic state prediction method based on the stochastic repetitive hill climbing method, and iteratively run the hill climbing method with the factors of weather environment, holidays and time of day as the influencing variables, and calculate the confidence level of directed edges to obtain a Bayesian network structure suitable for traffic state prediction and realize the inference prediction of traffic state [5, 7].

2 Bayesian Network

2.1 Overview

A Bayesian network is a binary group, i.e., $BN = (G, P)$, where $G = (X, E)$ denotes a Directed Acyclic Graph (DAG); $X = X_1, X_2, \dots, X_n$ is the set of nodes in the DAG, representing random variables; E is the set of directed edges, representing direct dependencies between variables; P is the additional probability distribution of variables. For the joint distribution $P(X_1, \dots, X_n)$ containing n variables in the network, according to the chain rule we have

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | \pi(X_i)) \quad (1)$$

This shows that Bayesian networks are a representation of the decomposition of joint probability distributions, which reduces the complexity of probabilistic models and also provides convenience for probabilistic inference.

2.2 Bayesian network modeling approach

The Bayesian network modeling process consists of two phases, network learning and inference learning. Network learning consists of two parts: structure learning and parameter learning. Structure learning is to find a network structure G_{best} that best matches the given training sample D . Parameter learning is to compute the probability distribution of the variables in the network based on the structure learning results using the given training data D' , i.e., to update the original prior probability distribution in the network.

2.2.1 Structural Learning

Due to the limitations of constructing Bayesian network structures relying only on domain knowledge, while searching for the optimal network structure is an NPhard problem, the effective use of data to automatically construct network structures has become the focus and difficulty in the field of Bayesian research. The existing research methods can be divided into two main categories: constraint-based and scoring-based.

Constraint-based structure learning algorithms are simple and fast, but the independence test is error-prone and directly affects the computation of the network structure, so it is difficult to guarantee the learning accuracy. The scoring-based structure learning algorithm defines a scoring function in advance and then searches for the highest rated network structure, so the scoring function and the search strategy are the main factors affecting the algorithm.

The most straightforward search strategy in the scoring-based structure learning algorithm is the exhaustive method, which calculates the score of each structure and then selects the structure with the highest score. In practice, this method is not feasible due to the large number of network structures. Hill-Climbing (HC) is a locally meritocratic search algorithm that selects the add-edge, subtract-edge, and trans-edge operators that improve the network score on a given initial structure, and iterates until the highest scoring network structure is obtained.

2.2.2 Parameter Learning

Parameter learning methods include maximum likelihood estimation and Bayesian estimation, which represent the classical and Bayesian schools of mathematical statistics, respectively.

The classical school considers the conditional probability $P(D|\Theta)$ of the training sample D as a function of Θ , and the maximum likelihood estimate of the parameter Θ is the Θ^* that corresponds to the maximum value of $P(D|\Theta)$, i.e.

The Bayesian school treats Θ as a random variable and calculates its posterior probability distribution. However, the integral operation required to calculate $P(\Theta|D)$ is usually difficult to implement, so an approximate method is usually used to solve the maximum a posteriori probability (Maximum A Posteriori, MAP), i.e., Bayesian MAP estimation.

Maximum likelihood estimation does not consider the effect of prior knowledge on the parameter to be estimated, while Bayesian estimation treats the parameter to be estimated as a random variable, which can be used with prior knowledge, and therefore Bayesian estimation is more reasonable.

2.2.3 Network Inference

Network inference refers to the use of Bayesian network structure and its probability distribution to calculate the posterior probabilities of variables under given conditions. The known variables are called evidence variables, denoted as E , and the values are denoted as e ; the variables for which the posterior probabilities need to be calculated are called query variables, denoted as Q , and their posterior probability distributions are $P(Q|E = e)$.

Bayesian network inference is also an NP-hard problem, and in recent years, some progress has been made in inference algorithms for specific types of Bayesian networks, such as variable elimination method and joint tree algorithm. Compared with the former, the joint tree algorithm is the most widely used exact

inference algorithm because it can improve the inference efficiency by sharing steps and has fast computation speed.

2.3 Network structure learning algorithm based on random repetition

This paper proposes to improve the above defects by Random Restart Hill-Climbing (RRHC), which is a local meritocratic algorithm and has the defect of easily falling into local optimum but not global optimum. Finally, the confidence level of each directed edge in the r network structures is calculated and the more frequent directed edges are selected by the confidence level threshold to form the final network structure.

3 Traffic state inference prediction model

3.1 Definition of traffic state inference prediction

Traffic state inference prediction refers to the construction of a Bayesian network with the traffic state identified by the traffic flow parameters as the outcome variable and the meteorological and temporal factors affecting the traffic state as the influencing variables; then the posterior probability of the traffic state under the given conditions is calculated and the maximum value is taken as the traffic state prediction result.

3.2 Traffic state inference prediction modeling process

According to the above Bayesian network modeling method, the traffic state prediction modeling process is divided into the following stages:

3.2.1 Data preprocessing and variable selection

The original data for a period of time is selected from a database storing road environment, meteorological conditions, traffic status information, etc.; the original data is pre-processed (including resampling at reset intervals, interpolation of missing data, discretization of continuous values, etc.); the influencing variables for constructing Bayesian networks are selected.

3.2.2 DAG random generation and directed edge blacklist setting

Set the number of randomly generated DAGs r ; set the blacklist of directed edges according to the interrelationship between variables and a priori knowledge, i.e., the DAGs obtained from each run of the hill-climbing algorithm do not contain any directed edges in the blacklist.

3.2.3 Stochastic repetitive Bayesian network learning

The scoring function is selected, and the hill-climbing algorithm is run iteratively with the randomly generated DAG as the initial structure; then the confidence of each directional edge and the confidence threshold are calculated for the obtained r DAGs, and then the optimal Bayesian network structure is obtained; finally, parameter learning is performed.

3.2.4 Inference prediction

The posterior probability of the query variable (traffic state) is calculated based on the values of the evidence variables, and the maximum value is taken as the inference prediction result of the traffic state.

3.2.5 Comprehensive evaluation of the model

The evaluation indexes are selected and the prediction results are evaluated comprehensively. If the evaluation result does not meet the prediction requirements, adjust the influencing variables and parameters in the model and repeat the above process until the model meets the evaluation requirements.

3.3 Model evaluation criteria

Based on the actual traffic status and Bayesian network prediction results, the confusion matrix is established as shown in Table 1.

Table 1: Confusion matrix of traffic status

Actual State	Predicted State				Total
	Free Flow c_f	Stable c_s	Crowded c_c	Congested c_j	
Free Flow c_f	$n_{f,f}$	$n_{f,s}$	$n_{f,c}$	$n_{f,j}$	N_f
Stable c_s	$n_{s,f}$	$n_{s,s}$	$n_{s,c}$	$n_{s,j}$	N_s
Crowded c_c	$n_{c,f}$	$n_{c,s}$	$n_{c,c}$	$n_{c,j}$	N_c
Congested c_j	$n_{j,f}$	$n_{j,s}$	$n_{j,c}$	$n_{j,j}$	N_j
Total	N_f	N_s	N_c	N_j	N

In Table 1, $n_{i,j}$ denote the number of samples in status category c predicted by Bayesian network as category c_j . The confusion matrix reflects the inferential prediction performance of Bayesian networks. The i -th row of the table reflects the Recall of category c_i , the j -th column reflects the Precision of category c_j , and the diagonal reflects the overall Accuracy. For a specific traffic state (e.g., blockage state c_j), the independent accuracy and recall are calculated as shown in equation (10) and equation (11), respectively, and the overall accuracy of the prediction is calculated as shown in equation (12).

Table 2: Directed edges whose confidence is greater than the confidence threshold

Directed edge	Confidence
Weather Conditions to Visibility	1.000
Weather Conditions to Hour	1.000
Day of Week to Holiday	1.000
Hour to Status	0.942
Holiday to Status	0.942
Weather Conditions to Day of Week	0.614

4 Case Studies

4.1 Data preprocessing and variable selection

The literature gives a basic method to apply traffic flow parameters for real-time traffic state identification. In this paper, we firstly apply the above method to identify the traffic flow parameters in December 2014 from the performance measurement system (PeMS) of the California Department of Transportation (Caltrans), and then build a traffic status database with the traffic flow parameters of the test station No. 717490 in Highway 101. After pre-processing, weather conditions, visibility, week, time of day, holidays, etc. were selected as the influencing variables.

4.2 Bayesian network learning results

The Bayesian Information Criterion (BIC) is chosen as the scoring function, and the hill climbing algorithm is run iteratively to obtain the 100 highest rated DAGs respectively. From Eq. (9), the final network structure is composed of the directed edges with confidence thresholds greater than 0.574 (see Table 2).

4.3 Analysis of traffic state prediction results

To summarize the above analysis, only two variables, such as time period and holiday, are selected as evidence variables, and the posterior probability of the query variable - traffic status is calculated by applying the joint tree algorithm, and the maximum value is taken as the traffic status prediction result. The confusion matrix was established by combining the traffic status under the actual conditions, and finally, various evaluation indexes of the prediction results were obtained (see Table 3).

As shown in Table 3, the overall accuracy of traffic status prediction exceeds 85% when only two variables such as time of day and holiday are selected for the detection station section, and the overall accuracy of traffic status prediction can reach 87.2% for three types of traffic status such as smooth, smooth and congested, which can provide forecast and early warning for monitoring the

Table 3: Evaluation results of traffic status prediction based on RRHC

Traffic State	Metric		
	Precision	Recall	Accuracy
Free Flow c_f	88.87	94.61	85.14
Stable c_s	81.63	75.33	
Crowded c_c	88.06	89.43	
Congested c_j	66.16	63.10	

Table 4: Evaluation results of traffic status prediction based on traditional HC

Traffic State	Metric		
	Precision	Recall	Accuracy
Free Flow c_f	86.49	89.83	79.39
Stable c_s	72.63	64.04	
Crowded c_c	81.88	87.41	
Congested c_j	59.41	49.65	

operation status of highway. The reason for the unsatisfactory prediction results is that the traffic congestion is affected not only by weekly, periodical and holiday factors, but also by occasional factors such as traffic accidents and bad weather, but the traffic congestion only accounts for 8.14% of the total sample of the network learning. The network learning results do not fully express the causality of the obstruction state.

From the comparison of Table 3 and Table 4, it can be seen that the traffic state prediction based on the random repetitive hill climbing method has significantly better evaluation indexes than the traditional hill climbing method, which makes up for the defect that the traditional hill climbing method is easy to fall into the local optimum.

5 Conclusion

In this paper, we proposed a traffic state prediction method based on the random repetitive hill climbing method by mining the basic data such as traffic state, weather environment and holiday information, and constructed a Bayesian network using this method to realize the inference prediction of traffic state. The method can provide an effective method and data support for highway operation status monitoring and early warning and decision analysis.

In this paper, the number of iterations, i.e., the number of DAG random generation, is specified in the random repetitive hill climbing process. Too many iterations will affect the learning speed of the network structure; too few iterations may still fall into local optimum. Therefore, the convergence of the

number of iterations needs to be further investigated. The traffic accident, bad weather and other incidental factors can have a significant impact on the traffic state, and how to improve the prediction results of the blockage state needs to be further investigated by filtering the historical data in a specific time period to build a Bayesian network.

References

- [1] N. Bekiaris-Liberis *et al.*, “Highway traffic state estimation with mixed connected and conventional vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3484–3497, 2016.
- [2] X. Sun *et al.*, “Highway traffic state estimation using improved mixture kalman filters for effective ramp metering control,” in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 6. IEEE, 2003, pp. 6333–6338.
- [3] D. A. Redelmeier *et al.*, “Traffic-law enforcement and risk of death from motor-vehicle crashes: case-crossover study,” *The Lancet*, vol. 361, no. 9376, pp. 2177–2182, 2003.
- [4] N. Friedman *et al.*, “Bayesian network classifiers,” *Machine learning*, vol. 29, no. 2, pp. 131–163, 1997.
- [5] J. J. Q. Yu, “Citywide traffic speed prediction: a geometric deep learning approach,” *Knowledge-Based Systems*, p. 106592, Nov. 2020.
- [6] S. Zhang *et al.*, “Bayesian deep learning for dynamic power system state prediction considering renewable energy uncertainty,” *Journal of Modern Power Systems and Clean Energy*, pp. 1–10, 2021.
- [7] J. J. Q. Yu *et al.*, “Real-time traffic speed estimation with graph convolutional generative autoencoder,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3940–3951, Oct. 2019.
- [8] —, “A social spider algorithm for global optimization,” *Applied Soft Computing*, vol. 30, pp. 614–627, May 2015.
- [9] C. Zhang *et al.*, “Spatial-temporal graph attention networks: a deep learning approach for traffic forecasting,” *IEEE Access*, vol. 7, pp. 166 246–166 256, 2019.
- [10] S. Sun *et al.*, “A bayesian network approach to traffic flow forecasting,” *IEEE Transactions on intelligent transportation systems*, vol. 7, no. 1, pp. 124–132, 2006.
- [11] J. Sun *et al.*, “A dynamic bayesian network model for real-time crash prediction using traffic speed conditions data,” *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 176–186, 2015.
- [12] J. J. Q. Yu *et al.*, “Intelligent time-adaptive transient stability assessment system,” *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1049–1058, Jan. 2018.
- [13] —, “Intelligent fault detection scheme for microgrids with wavelet-based deep neural networks,” *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1694–1703, Mar. 2019.
- [14] —, “Online false data injection attack detection with wavelet transform and deep neural networks,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3271–3280, Jul. 2018.
- [15] —, “A social spider algorithm for solving the non-convex economic load dispatch problem,” *Neurocomputing*, vol. 171, pp. 955–965, Jan. 2016.
- [16] Y. Liu *et al.*, “Privacy-preserving traffic flow prediction: a federated learning approach,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, Aug. 2020.
- [17] F. Zong *et al.*, “Prediction for traffic accident severity: comparing the bayesian network and regression models,” *Mathematical Problems in Engineering*, vol. 2013, 2013.

- [18] J. J. Q. Yu *et al.*, “Online vehicle routing with neural combinatorial optimization and deep reinforcement learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3806–3817, Oct. 2019.
- [19] —, “Autonomous vehicle logistic system: Joint routing and charging strategy,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 7, pp. 2175–2187, Jul. 2018.
- [20] A. Miglani *et al.*, “Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges,” *Vehicular Communications*, vol. 20, p. 100184, 2019.
- [21] Y. Lv *et al.*, “Traffic flow prediction with big data: a deep learning approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.
- [22] J. J. Q. Yu, “Two-stage request scheduling for autonomous vehicle logistic system,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1917–1929, May 2019.
- [23] J. J. Q. Yu *et al.*, “Sensor deployment for air pollution monitoring using public transportation system,” in *Proc. IEEE Congress of Evolutionary Computation*. Brisbane, Australia: IEEE, 2012, pp. 1–7.
- [24] L. Zhang *et al.*, “An improved k-nearest neighbor model for short-term traffic flow prediction,” *Procedia-Social and Behavioral Sciences*, vol. 96, pp. 653–662, 2013.
- [25] J. J. Q. Yu *et al.*, “Double auction-based pricing mechanism for autonomous vehicle public transportation system,” *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 2, pp. 151–162, Jun. 2018.
- [26] J. J. Q. Yu, “Semi-supervised deep ensemble learning for travel mode identification,” *Transportation Research Part C: Emerging Technologies*, vol. 112, pp. 120–135, Mar. 2020.
- [27] Y. Wu *et al.*, “A hybrid deep learning based traffic flow prediction method and its understanding,” *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 166–180, 2018.
- [28] J. J. Q. Yu, “Sybil attack identification for crowdsourced navigation: a self-supervised deep learning approach,” *IEEE Transactions on Intelligent Transportation Systems*, 2020, in press.
- [29] N. G. Polson *et al.*, “Deep learning for short-term traffic flow prediction,” *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.
- [30] H. Yin *et al.*, “Urban traffic flow prediction using a fuzzy-neural approach,” *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 2, pp. 85–98, 2002.
- [31] J. J. Q. Yu *et al.*, “Synchrophasor recovery and prediction: a graph-based deep learning approach,” *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7348–7359, Oct. 2019.