Expressway traffic risk intelligent early warning method based on Bayesian network

Lai Kuntao[1,*]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044 P. R. China

Abstract. Expressway traffic hazards often evolve into traffic accidents. Because of the potential traffic risks and system complexity, it is difficult to deal with the real-time expressway traffic risk early warning problem by relying solely on the experience of decision makers and scattered monitoring data. Therefore, it is necessary to study the theory and method of expressway traffic risk early warning by means of data-driven decision-making approach, that is, relying on traffic big data technology to construct a holographic view of expressway traffic status for decision makers, excavating the anomalies hidden behind the data sources, and characterizing traffic accidents. This article focuses on expressway traffic risk intelligent early assessment, using the MATLAB toolbox BNT to establish a Bayesian network for expressway traffic for assessing the risk, discussing the validity and interpretability of the model. The accuracy of the training set and test set is about 0.8902 and 0.8874, respectively, which verifies the model is acceptable and valid. The innovation of this paper is to deal with the problem of expressway traffic risk early warning based on the data-driven perspective, and focuses on the interpretability of the model, giving the expressway decision makers adequate warning information.

1 Introduction

Traffic risk warning can achieve the purpose of eliminating traffic risks ahead of time by judging the precursors of traffic accidents; integrating multi-source data to improve the decision-making ability of expressway managers; guiding drivers to pay attention to risk sources and risk levels and coping with traffic accidents.

Aiming at the traffic risk intelligent warning of expressway, the general research idea is to first obtain the traffic operation status and traffic accident database of the expressway over the years, then extract the main factors affecting the expressway, and then select the appropriate mathematical model to establish the expressway traffic accident and forecast factor. The relationship model is finally applied to the real-time intelligent decision-making scheme of managers.

There are many factors leading to accidents and intertwined factors, and there is a complex relationship with traffic accidents, it is necessary to select subsets from many predictors sets to meet the requirements that variables can be obtained, and factors can maximize the prediction effect. Hossian introduced a new variable congestion coefficient (CI, Congestion Index) to eliminate the impact of speed on road conditions [1]. Sun only used the driving speed as a model variable, effectively solving the over-fitting and generalization of the model caused by redundant variables. The problem of not capable ability can reduce the calculation and learning complexity of the model and improve the running ability [2]. Li divides the overall traffic state into free flow, blocking flow, evacuation flow and assembly flow according to the upstream and downstream speeds. The conclusion is that the speed standard deviation variable has greater influence in the free flow traffic state, and the velocity in the assembly flow [3].

Prediction model can be roughly divided into statistical model and machine learning model. The former is generally classified tree, logistic regression, and other models. The latter has SVM, random forest, neural network, Bayesian Network, and other models. Yu believes that in the previous research, the Logistic regression model has a collinearity problem, the neural network model has the disadvantage of over-fitting, and the model is ignored. The generalization ability only focuses on parameter estimation [4]. Basso found that the predecessor training model chose to use balanced data sets, which did not meet the actual situation of small accidents in traffic events, and established SVM and Logistic model [5]. In the process, SMOTE (Synthesis of a few oversampling techniques) techniques can be used to train unbalanced data sets. Sun compared logistic regression, Bayesian network, Naive Bayes, K-nearest, backward propagation neural network and support vector machine to predict the occurrence of traffic accidents. The support vector machine achieves a maximum accuracy of nearly 80% [6].

However, there is still room for improvement in the research status of expressway traffic risk warning theory and method: the early warning model cannot be selected, and the early warning model is more biased than the

[*] Corresponding author: 18120815@bjtu.edu.cn

forecasting accuracy. Most of the intelligent algorithms with "black box" effect are adopted. The early warning model is regarded as a classification problem. Although the prediction result of the model is considerable, it lacks the explanatory nature of the early warning model, that is, it answers the "risk", but does not inform the manager "what is the risk"; the classification of traffic risk is subjective. Most of the traffic risk grades are classified using analytic hierarchy process, fuzzy mathematics evaluation and other subjective algorithms. There are rigorous problems and suspected fraud. This cannot follow the "data-driven" idea to divide traffic risk grades scientifically and reasonably. In addition, it is not practical to build a model with too many variables. Many studies have applied some unrelated variables to the early warning model, and they have fallen into the strange circle of "the more accurate the model is," and some of the variables are for expressway traffic safety. Management and control are difficult to change, such as social population and travel variables, and it is meaningless to discuss these variables in an early warning model.

2 Materials and methods

2.1 Study design

Aiming at the traffic risk assessment problem of the I-5 interstate expressway in San Diego, USA, a Bayesian network model with input of important characteristics related to traffic accidents and output with traffic risk value is established. The characteristic variables are extracted from the environment, road, and vehicle. The correlation matrix of variables is used to observe the correlation of characteristic variables to traffic accidents, and the causal map of Bayesian network is constructed. The maximum likelihood estimation method is used to implement the parameter learning of Bayesian network, and finally the validity of the model is verified.

2.2 Bayesian network

The Bayesian network is a directed graph in which each node is labelled with quantitative probability information, which is fully described below.

(1) Each node corresponds to a random variable, which can be discrete or continuous.

(2) A set of directed edges or arrows connect the node pairs. If there is an arrow pointing from the node to the node, it is called a parent node.

(3) Each node has a conditional probability distribution that quantifies the impact of its parent on that node.

Bayesian networks can be understood from both qualitative and quantitative levels. At the qualitative level, it uses a directed acyclic graph to describe the dependencies and independent relationships between variables; at the quantitative level, it uses conditional probability distributions to characterize the dependencies of variables on their parent nodes. Semantically, Bayesian networks are a representation of joint probability distribution decomposition. Specifically, assuming that the variables in the network are $X_1, X_2, \dots, X_n$, then the probability distributions attached to the variables are commensurate to obtain a joint distribution,

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

In practical applications, people often use causality to determine the structure of Bayesian networks and determine the order of variables. Causal relationships often make the network structure simple and the probability distribution is easy to evaluate. In a Bayesian network built using causality, the edges between variables represent a causal relationship rather than a simple probability dependency. Such a Bayesian network is called a Bayesian causal network. In Bayesian causal networks, in addition to probabilistic reasoning, reasoning about the consequences of interventions and false reasoning can also be performed.

3 Results analysis and discussion

3.1 Model Features

There are 667 accident data, and there are 2668 corresponding non-accident data. 14 variables are extracted from the three perspectives of environment, road, and vehicle. The specific variables and corresponding symbols are shown in Table 1.

Bayesian networks can deal with discrete variables and continuous variables, but continuous variables have constraints on their distribution. The general practice is to discretize continuous variables, that is, data binning. The quality of data binning determines the upper limit of model accuracy. Some simple binning methods are classified according to the conventions of the real world. In addition, the limit values given by the relevant specifications are binned. However, some variables do not have ready-made classification methods, including congestion coefficient, traffic saturation and time occupancy. In this case, mathematical methods are needed for classification. At the same time, these three variables will be related to the traffic risk value directly, and the binning effect will largely affect the classification results of the Bayesian network. Because the training data is not classified, the process of data binning is unsupervised clustering. Common unsupervised bins are equidistant bins and equal-frequency bins. The equidistant binning is used here. This method is simple to operate and can truly reflect the distribution of data. To improve the accuracy, the number of bins is increased as much as possible without affecting the generalization ability of the model. The specific variables are shown in Table 1.

Table 1. Variable set

Variable	Value
Circumstance	
Time	1. day 2. night
Working day	1. working day 2. non-working day
Weather	1. sunny day 2. non-sunny day
Wind speed	1. Wind speed is not more than 30MPH 2. Wind speed is greater than 30MPH

Visibility	1. Visibility is less than 50 meters 2. Visibility is less than 100 meters and not less than 50 meters 3. Visibility is less than 200 meters and not less than 100 meters 4. Visibility is greater than 200 meters
Road	
Round curve radius	1. Radius is not more than 0.6 mile 2. Radius is greater than 0.6 mile and no more than 6 mile 3. Radius is greater than 6 mile
Road material	1. Cement or asphalt 2. Others
Number of lanes	1. One-way six lanes 2. One-way five lanes 3. One-way four lanes
Road width	1.Width 48meter 2.Width 60meter 3.Width 72meter
Dry state	1.Dry 2.Wet
Vehicle	
Upstream & downstream status	1.FF 2.BN 3.BQ 4.CT
Congested Index	15 levels
Traffic saturation	15 levels
Average occupancy	10 levels

The value of each variable will be explained below.

(1) Time. The morning time range is defined as 6:00 to 18:00, and the night-time range is defined as 18:00 to 6:00 the next day.

(2) Working day. The working day range is defined as Monday through Friday, and the non-working day range is defined as Saturday through Sunday.

(3) Weather. Sunny day is defined as cloud coverage less than 40% and no rain or snow, otherwise it is not sunny.

(4) Wind speed. According to the wind level division method, 30MPH (about 50KM/H) is a classification level of fresh wind and strong wind and has lateral force on vehicle driving.

(5) Visibility. According to the implementation regulations of the Road Traffic Safety Law, when the visibility is less than 200 meters, the speed does not exceed 60 MPH. When the visibility is less than 100 meters, the speed does not exceed 40 MPH. When the visibility is less than 50 meters, the speed does not exceed 20 MPH and leave the expressway as soon as possible.

(6) Round curve radius. According to expressway technical specifications, when the design speed is 75MPH (about 120KM/H), it is more dangerous to have a radius of less than 0.6mile (about 1000m). When the radius is greater than 6 miles (about 10,000 meters), it is close to a straight line.

(7) Pavement material. There are two types of pavement materials: cement or asphalt and other materials.

(8) Number of lanes. There are three types of lanes: one-way six lanes, one-way five lanes, and one-way four lanes.

(9) Road width. There are three types of road widths: 48m, 60m and 72m.

(10) Dry state. Whether there is current rainfall or not, to determine whether the road surface is dry.

(11) Upstream and downstream traffic status. Judging the average speed of the upstream and downstream detectors, the speed of 50MPH is the basis for the classification of traffic steady flow and unsteady flow. According to the traffic stability of upstream and downstream, the four traffic states are: free flow, forward propagation, backward propagation, Blocking flow.

(12) Congestion index. The speed of speed has a direct causal relationship with traffic accidents. Since the road geometry of the location where the station is located is different, simply using the average speed as an indicator is not comparable. Therefore, the Congested Index should be introduced, which is defined as

$$CI = \frac{FFS - V}{FFS} \tag{2}$$

CI represents the congestion index, FFS represents the free-flow velocity.

(13) Traffic saturation. The magnitude of traffic saturation on the road has a direct impact on the occurrence of traffic accidents. Traffic saturation cannot be used directly to avoid the influence of road geometry. Therefore, pick traffic saturation as an indicator, the calculation formula is

$$V/C = \frac{\text{Volume (veh }/h)}{\text{Capacity (veh }/h)} = \frac{\text{Volume }(veh/5min) \cdot 12}{\text{Capacity }(veh/h/ln) \cdot \text{ lane}} \tag{3}$$

Capacity $(veh/h/ln)$ represents the maximum capacity of each lane of the expressway.

(14) Average occupancy. To characterize the traffic congestion of the road, the road occupancy rate is used. When the vehicle is within the effective range of the detector, the detector can be kept in the on state, and t_i the duration of the vehicle passing through the detector can be measured by the timing device, and the calculation formula is

$$R_t = \frac{\sum_{i=1}^{n} t_i}{t} \times 100\% \tag{4}$$

R_t represents the time share, t represents the total observation time.

If the characteristic variables of the original data directly as an indicator of traffic risk is used, high-dimensional data sets will be very tricky. Too high a dimension can make most learning algorithms inefficient or even impossible to get the result. Therefore, dimensionality reduction is a necessary task before establishing a system. Depending on the degree of correlation between variables, the relationship between variables with small correlation coefficients is not considered. Draw a matrix of correlation coefficients between the 14 variables, as shown in Figure 1. It is intuitively found from the figure that there are many uncorrelated correlation coefficients, which can greatly simplify the Bayesian network structure constructed later, and also help to improve the generalization ability and robustness of the model.

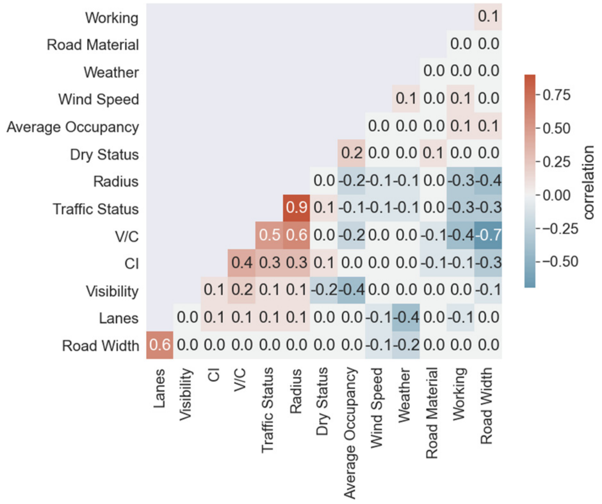

Fig. 1. Correlation coefficient matrix

3.2 Model establishment

According to the correlation matrix diagram, four variables with little correlation are filtered to ensure that the child nodes of the same parent node have little correlation. Finally, it is determined that the nodes participating in the formation of the Bayesian network have 8 discrete variables, namely weather conditions, visibility, time, road width, traffic saturation, upstream and downstream traffic state, radius of the curve and traffic risk.

The vehicle's characteristic variables include traffic saturation and upstream and downstream traffic conditions are linked to traffic risk values, which characterizes the traffic operation and represents the driver's driving status. The two describe the traffic flow state from two sides, one from the state of the accident point itself, and the other from the traffic state of the upstream and downstream of the accident point. In addition, the occurrence of traffic accidents is the effect of vehicles on the road, then the advantages and disadvantages of traffic flow factors will directly determine the value of traffic risk.

Visibility and weather conditions are linked because of the strong correlation between the two, that is, poor visibility in poor weather conditions. The time and road width are respectively connected to the traffic saturation because the traffic saturation has obvious time-varying laws, and the road width affects the traffic saturation. Other conditions are better and are not considered in this model. In summary, the network structure of the Bayesian network is shown in Figure 2.

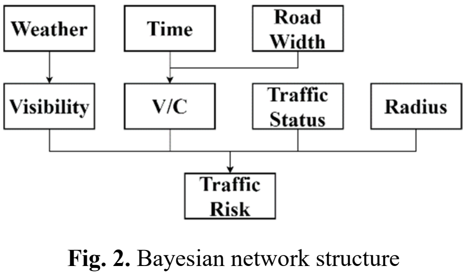

Fig. 2. Bayesian network structure

Bayesian network parameter estimation is an estimation of the Condition Probability Table (CPT) between variables. Using MATLAB's toolbox BNT, the maximum likelihood estimation method is used to estimate the parameters, as shown in Figure 3.

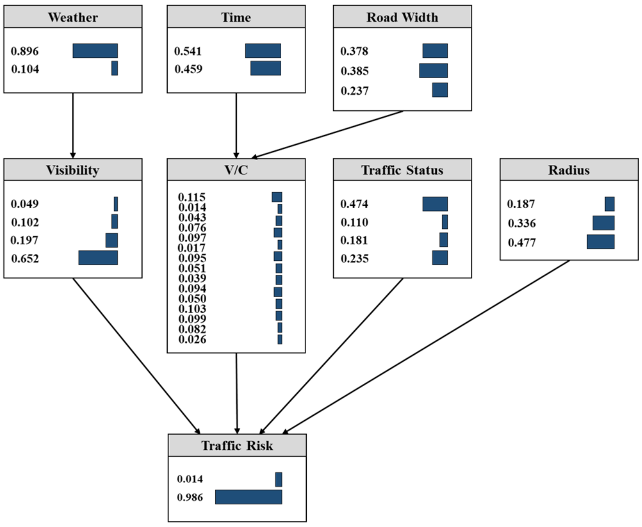

Fig. 3. Bayesian network conditional probability distribution

3.4 Model verification

In general, Bayesian network is a machine learning method, and the output is judged by cross-validation method to judge the reliability of the result. In the calculation, 5 training sets are randomly established, corresponding to 5 test sets, wherein the number of samples in the test set is 0.3 of the total number of samples.

In addition to using accuracy as an indicator for evaluating the classification model, there is a significant difference between the number of accident samples and the number of non-accident samples. The accuracy of one cannot fully reflect the accuracy of the model. Since a traffic accident is a small probability event, if the model output is a non-accident warning result, the accuracy rate will be high, which will cover up the error of unpredictable accident. Therefore, two additional indicators need to be set: accuracy rate and recall rate. A confusion matrix is assumed to summarize the classification results of the Bayesian network model, which is defined as shown in Figure 4.

True Positives **(TPs)** Prediction: Crash Correct: Crash	False Positives **(FPs)** Prediction: Crash Correct: Non-crash
False Negatives **(FNs)** Prediction: Non-crash Correct: Crash	True Negatives **(TNs)** Prediction: Non-crash Correct: Non-crash

Fig. 4. Confusion matrix definition

Then, the reliability indicators (Accuracy), Precision (Precision), and Recall (Recall) are defined as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

There is a trade-off between accuracy and recall, which means that increasing the accuracy usually reduces the recall rate and vice versa. In order to determine the threshold of the classification, the training set is used to plot the accuracy of a series of classification thresholds and the recall rate, as shown in Figure 5.

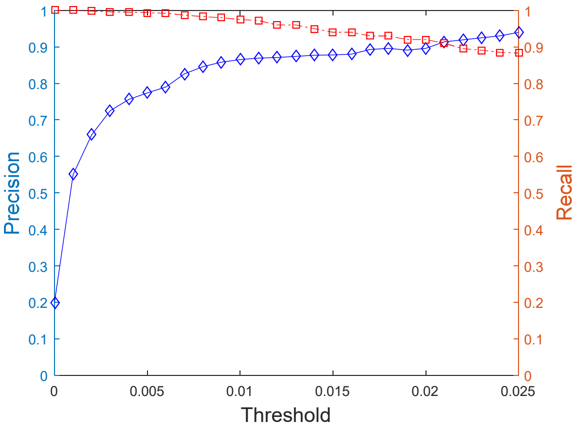

Fig. 5. The relationship between precision and recall and threshold

It can be seen from the figure that as the classification threshold increases, the false positives decrease and the exact value becomes larger; when the false negatives increase, the recall rate becomes smaller. At about 0.02, the exact value is equal to the recall rate, so the maximum accuracy can be achieved when the classification threshold is set to 0.02. The calculated values are shown in Table 2.

Table 2. Reliability index

Reliability index	Value
Accuracy	
Training set	0.8902
Test set	0.8874
Precision	
Training set	0.8824
Test set	0.8755
Recall	
Training set	0.8874
Test set	0.8791

The results show that the index values of the training set, and the test set are close, indicating that the Bayesian network model has better generalization ability and stable model accuracy. The higher accuracy value indicates that the classification effect of the model is better, that is, the expressway traffic risk warning result is good. The high precision and recall rate values indicate that the probability of marking an accident event as a non-accident event is small and has a good application effect.

4 Conclusion

This paper focuses on the theory and method of expressway traffic risk warning. Overall, the data drive drives the traffic risks ahead of time and assists the expressway safety department in making traffic safety management and control measures. The full text emphasizes the interpretability of the model, which is conducive to providing a wealth of evidence of traffic accidents, not only the more accurate real-time detection of traffic safety anomalies, but also the risk value and traffic risk rating of each factor. Managers can identify potential traffic hazards in time and eliminate existing traffic risks in advance.

The traffic flow data studied in this paper is collected from a single induction coil. The number of traffic flow characteristic variables is not as high as that of radar monitoring stations and dual induction coils. For example, the standard deviation of the location speed and the coefficient of variation and the mixing rate of the cart. If the traffic flow characteristic variables can be increased, the discussion of the correlation between the traffic accidents will be more profound.

References

1. Hossain, M. & Muromachi, Y. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accid. Anal. Prev.* **45**, 373–381 (2012).
2. Sun, J. & Sun, J. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transp. Res. Part C Emerg. Technol.* **54**, 176–186 (2015).
3. Li, Z., Wang, W., Chen, R., Liu, P. & Xu, C. Evaluation of the Impacts of Speed Variation on Freeway Traffic Collisions in Various Traffic States. *Traffic Inj. Prev.* **14**, 861–866 (2013).
4. Yu, R. & Abdel-Aty, M. Utilizing support vector machine in real-time crash risk evaluation. *Accid. Anal. Prev.* **51**, 252–259 (2013).
5. Basso, F., Basso, L. J., Bravo, F. & Pezoa, R. Real-time crash prediction in an urban expressway using disaggregated data. *Transp. Res. Part C Emerg. Technol.* **86**, 202–219 (2018).
6. Sun, J., Sun, J. & Chen, P. Use of Support Vector Machine Models for Real-Time Prediction of Crash Risk on Urban Expressways. *Transp. Res. Rec.* **2432**, 91–98 (2014).