Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor

Li Kuang, Han Yan, Yujia Zhu, Shenmei Tu & Xiaoliang Fan

To cite this article: Li Kuang, Han Yan, Yujia Zhu, Shenmei Tu & Xiaoliang Fan (2019) Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor, Journal of Intelligent Transportation Systems, 23:2, 161-174, DOI: 10.1080/15472450.2018.1536978

To link to this article: https://doi.org/10.1080/15472450.2018.1536978

Published online: 20 Feb 2019.

Submit your article to this journal

Article views: 33

View Crossmark data

Citing articles: 1 View citing articles
Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor

Li Kuanga, Han Yan, Yujia Zhub, Shenmei Tua, and Xiaoliang Fanb,c

aSchool of Software, Central South University, Changsha, China; bFujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China; cDigital Fujian Institute of Urban Traffic Big Data Research, Xiamen University, Xiamen, China

ABSTRACT

With the development of urbanization, road congestion has become increasingly serious, and an important cause is the traffic accidents. In this article, we aim to predict the duration of traffic accidents given a set of historical records and the feature of the new accident, which can be collected from the vehicle sensors, in order to help guide the congestion and restore the road. Existing work on predicting the duration of accidents seldom consider the imbalance of samples, the interaction of attributes, and the cost-sensitive problem sufficiently. Therefore, in this article, we propose a two-level model, which consists of a cost-sensitive Bayesian network and a weighted K-nearest neighbor model, to predict the duration of accidents. After data preprocessing and variance analysis on the traffic accident data of Xiamen City in 2015, the model uses some important discrete attributes for classification, and then utilizes the remaining attributes for K-nearest neighbor regression prediction. The experiment results show that our proposed approach to predicting the duration of accidents achieves higher accuracy compared with classical models.

Introduction

In recent years, intelligent transportation provides all-round support for transportation planning, management, transportation and public travel, by effectively integrating sensor technology, network communication technology, data processing technology and business application (Fan, Khattak, & Shay, 2007; Ran, Jin, Boyce, Qiu, & Cheng, 2012). It not only promotes the precision and intelligence of traffic management, but also improves the operational efficiency and service levels of transportation systems. The governance of urban traffic congestion is an important research topic in the field of intelligent transportation. Among them, traffic accidents are an important cause of road congestion and they often happen occasionally, so how to predict the duration of traffic accidents from occurrence to release quickly and accurately, so as to guide the congestion and restore the road, has become a significant and challenging research problem.

The rich historical data collected by various kinds of vehicle sensors provide the possibility of predicting the duration of traffic accidents. Ring coil can detect road traffic conditions, such as traffic flow, speed, occupancy, and vehicle length. The ultrasonic sensor utilizes the influence of the shape of vehicles on the ultrasonic wave front to detect whether the vehicle is approaching or moving away. Video vehicle detectors can be highly efficient wide-area video surveillance and real-time collection of various traffic parameters. Piezoelectric sensors are useful for dynamic weighing, vehicle classification statistics, speed detection, and parking area monitoring. We can extract the start time, the end time, the location of the accident, the number of vehicles involved in the accident and the description of the accident from the original data obtained by sensors, and construct the prediction model with these characteristic data to predict the duration of traffic accidents.

In the related work on predicting the duration of traffic accidents, scholars initially use single models such as decision tree, neural network, and multiple linear regression models. In order to improve the prediction accuracy, some scholars consider the advantages of combining multiple models, in which traffic accidents are classified from the perspective of clustering first, and then regression algorithms for each type of accidents are used to predict the exact
values. In addition, there are some studies aiming at solving the heterogeneity of traffic accidents, the confidence interval of predicted values, and the probability of occurrence of the second accident.

However, there are still some shortcomings in existing work: (1) the number of traffic accident duration samples has been proved to be unbalanced, but few studies have solved the problem. (2) In the process of modeling with the characteristic attributes, the interaction effects between the attributes are not fully considered, which may affect the prediction accuracy of duration values. (3) The prediction of traffic accident duration is a cost-sensitive problem in practice. For example, if a long-lasting accident is estimated incorrectly, vehicles will be misled to select a wrong route, thereby increasing the congestion and road pressure, while if a short-lasting accident is estimated incorrectly, vehicles will plan other routes, only slightly increasing the car energy consumption. However, there is little work on the introduction of cost-sensitive issues into the prediction of traffic accident duration.

Aiming at addressing the shortcomings of existing work, in this article, we propose a method of predicting the traffic accident duration based on the combination of cost-sensitive Bayesian network and weighted K-nearest neighbor (KNN). First, we make a two-factor variance analysis of discrete features by collecting the external parameters of accidents, and constructing a network topology diagram among features based on MMPC(Max-Min Parents and Children) and K2(a heuristic search algorithm). Then, a Bayesian network model is constructed by introducing a cost-sensitive function, which divides the duration of the accident into two classes, that is, more than 30 min and less than 30 min. Finally, after determining the class label, KNN regression based prediction is performed using the continuous characteristic parameters to obtain the predicted duration value.

KNN regression model is a nonparametric lazy learning algorithm and works well in practice. It first finds similar K neighbors by defining the similarity distance of input features, and then predict the target value based on the records of K neighbors. In the application of predicting accidents duration, we find that under similar context including time and location, the duration of accidents are similar. Therefore, it is natural to employ KNN regression to capture the similarity of durations under similar context.

The remainder of the article is organized as follows: Related work section discusses the related work. Predicting the accident duration section defines the problem to be solved and illustrates the detailed solution to predicting the duration of traffic accidents. Experiment section presents the experimental verification design and result analysis. Finally, Conclusion section gives conclusions and future work.

Related work

In recent year, the problem of predicting the duration of traffic accidents has attracted wide attention of scholars. Related work on the research issue focus on the following aspects: (1) investigating the statistical characteristics of traffic accident data; (2) exploring the influence factors which affect the length of accident; (3) constructing a proper prediction model.

Usually it is a first step to investigate the statistical characteristics of the accident data before building the prediction model, such as the distribution of accident duration. Such statistical characteristic of data is important to the choice of prediction model, as well as the choice of data discretion. Some researchers work on the distributions of accident duration and find different distributions in different dataset, such as logarithmic normal distribution (Garib, Radwan, & Al-Deek, 1997; Giuliano, 1989; Golob, Recker, & Leonard,1987), log-logical distribution (Jones, Janssen, & Mannering, 1991), Weibull distribution(Hojati, Ferreira, Washington, & Charles, 2013; Nam & Mannering, 2000); Gamma distribution(Li & Guo, 2015), generalized distribution, etc. For example, Golob et al. (1987) proposed to use Markov to test the duration of a truck traffic accident and find the durations follow the logarithmic normal distribution, and the significant levels are from 0.31 to 0.99. Subsequently, Giuliano (1989) and Garib et al. (1997) also proved that the durations of various kinds of traffic accidents subject to logarithmic normal distribution. Jones et al. (1991) conducted appropriate statistical analysis through the frequency and duration of accidents on the Seattle expressway, and found that the accident data follow log-logical distribution. Nam and Mannering (2000) divided the duration of the accident into four stages, due to the mutual influence between every two stages, different models are used for each stage, and each stage corresponds to log-normal distribution, double logarithmic distribution, and Weibull distribution, respectively.

Feature recognition and construction is also important to model building. Many researchers work on exploring the features that affects the accident duration. The main features that they have found include: accident type (rear-end collision, turn over, vehicle breakdown, etc.) (Nam & Mannering, 2000) and accident severity; geographical location
characteristics (longitude and latitude, cross road information) (Li & Guo, 2015); the number and type of vehicles involved (trucks, buses, cars, etc.) (Chung, 2010); road surface condition (dry, wet, snowy, and icy conditions) (Chung, 2010), reporter type (independent drivers, Freeway Service Patrol, Traffic Information Service Provider, and Freeway Information Center) (Chung, 2010); the lane type (left, middle, right, emergency lane) (Kang & Fang, 2011); weather (sunshine, cloudy, rain, snow, fog) (Alkaabi, Dissanayake, & Bird, 2011); time characteristics (non-peak days, peak days, nights, weekdays, weekends) (Hojati et al., 2013); incident characteristic (severity, type, injury, medical requirements, etc.) and infrastructure characteristics (roadway shoulder availability) (Hojati, Ferreira, Washington, Charles, & Shobeirinejad, 2014). For example, Chung (2010) studied the duration of traffic accidents on freeway in Korea, and they combined time, location, accident type, involved vehicle type, accident severity, road surface condition, and reporter type to build prediction model. Wang, Cong, and Qiao (2013) used the combination of weather, time when policeman arrived, accident type, and lanes type to model the accident dataset of freeway in Zhejiang province. Being different from the freeway, there are some differences in factors that affect the accident duration in urban roads. And it is not definite to employ more features to get better results. In the study of Alkaabi et al. (2011), they presents the results of investigating the effects of features on the accident clearance time with emphasize of accelerated failure time (AFT) metric, and before modeling, they carry out correlation analysis, so only useful features are kept.

The main challenge of predicting accident duration focuses on constructing a proper prediction model. Researchers first tried to apply various kinds of single models in the field, and the models can be roughly divided into four kinds: regression model (Khattak, Schofer, & Wang, 1995; Valenti, Lelli, & Cucina, 2010; Wang, Cong, et al., 2013; Wu, Chen, Zheng, et al., 2011), tree classification model (Boyles, Fajardo, & Waller, 2007; Kim, Chang, Rochon, et al., 2008; Pan, Wang, Zhan, & Deng, 2018; Zhan, Gan, & Hadi, 2011), artificial neural network (Park, Haghani, & Zhang, 2016; Valenti et al., 2010; Vlahogianni & Karlaftis, 2013; Wei & Lee, 2007), and statistical modeling (Giuliano, 1989; Golob et al., 1987; Hojati et al., 2013; Jones et al., 1991).

1. Regression model: Khattak et al. (1995) proposed to use truncated regression models to verify a series of hypotheses on influence factors first, and then develop a time sequential methodology to predict the incident durations. Valenti et al. (2010) proposed to apply several models to incident duration prediction and compare their performance, include multiple linear regression, support/relevance vector machine and K-nearest neighbor.

2. Tree classification model: By comparing different models including simple linear regression model and two nonparametric regression models, Koppelman, Sethi, and Ivan (1994) established the accident duration model based on decision tree method. Based on the Classification and Regression Tree (CART) and the findings from preliminary analysis of data set. Kim et al. (2008) has redesigned a classification tree named Rule-Based Tree Model (RBTM) to identify variables influencing the incident duration and estimate incident duration. Zhan et al. (2011) proposed to utilize M5P tree algorithm for predicting the incident duration.

3. Artificial neural network: Wei and Lee (2007) creates an adaptive procedure for sequential forecasting of incident duration, which includes two adaptive artificial neural network-based models as well as the data fusion techniques. Vlahogianni and Karlaftis (2013) think that incident duration data is incomplete and inaccurate, and they address the problem of incident duration prediction from the survival analysis perspective using advanced artificial intelligent techniques.

4. Statistical models: Researchers also try to employ probability distribution model (Giuliano, 1989; Golob et al., 1987), and discrete selection model (Jones et al., 1991) to predict accident duration. For example, Hojati et al. (2013) proposed a hazard-based duration modeling approach to model incident duration as a function of a variety of factors that influence traffic incident duration.

In order to address the limited ability of single model and further improve the prediction accuracy, some scholars considered combining the advantages of multiple models. Li and Guo (2015) proposed a mixture model which uses the multinominal logistic model and parametric hazard-based model to assess the influence of covariates on the probability of clearance methods and on the duration of the incident. Ghosh, Asif, and Dauwels (2016) propose Bayesian Support Vector Regression (BSVR), which gives error bars as the measurement of uncertainty along with the predicted duration of incidents. They also evaluate sensitivity and specificity for different error tolerance
limit to assess the performance of BSVR. In the research study of He, Kamarianakis, Jintanakul, and Wynter (2013), they proposed a hybrid tree-based quantile regression method and quantification of the effects of various incident and traffic characteristics that determine duration. They show that hybrid tree-based quantile regression incorporates the merits of both quantile regression modeling and tree-structured modeling: robustness to outliers, simple interpretation, flexibility in combining categorical covariates, and capturing nonlinear associations. Lin, Wang, and Sadek (2016) proposes a novel approach for accident duration prediction, which improves on the original M5P tree algorithm through the construction of a M5P-Hazard-Based Duration Model (HBDM) model, in which the leaves of the M5P tree model are HBDMs instead of linear regression models. And the proposed M5P-HBDM managed to identify more significant and meaningful variables than either M5P or HBDMs. The multimodels indeed improve the accuracy of duration prediction, but the effects of interaction between attributes are not fully taken into consideration in the multimodels. As well, there is little work on introducing cost-sensitive issues into the prediction of accident duration.

Cost-sensitive learning is a kind of problems in which the cost of missing a target is much higher than that of a false-positive, and classifiers with respect to losses are designed to weigh certain types of errors more heavily than others. Cost-sensitive learning has been imported into the problem of fraud detection, medical diagnosis, or object detection in computer vision (Deng & Chen, 2015; Kuang et al., 2018; Liao et al., 2017; Yang, Wang, Mi, Lin, & Cai, 2009; Zeng et al., 2018). In the research of Liu, Zhang, Zhang, and Wang (2011), cost-sensitivity is also introduced to analyze uncertain data in traffic flow prediction. However, cost-sensitive issue has not been fully addressed in existing solutions to accident duration prediction.

Compared with the existing related work, our main innovations include: (1) most of the existing work just use a single model to predict the duration of accidents, either classification model or regression model, however, in this article, based on a thorough analysis on data, we explore to classify the accident duration into two categories first, that is, less or larger than 30 min, and then predict the detailed value with a regression model. (2) Most of the existing work do not fully consider the interaction effect of influence factors, assuming that the factors are independent with each other. However, they are not independent, for example, a bad weather may lead to a serious accident. Therefore, in this article, we explore the interaction between influence factors and employ Bayesian network to model their interaction relation. (3) We import cost-sensitive problem into the prediction of accident duration. If an accident with long duration is predicted to be a short one, more vehicles will be guided to choose the road, and it will lead to a more serious traffic congestion. Therefore, in this article, we propose a cost-sensitive Bayesian network for duration classification, so as to guarantee that a wrong classification on accident with long duration to a short one would get a heavy penalty.

### Predicting the accident duration

#### Problem definition

Given the data set \( R = \{r_1, r_2, \ldots, r_n\} \), where \( r_i \) denotes the \( i \)th traffic accident history, \( r_i \) can be represented as a 6-tuple, \( r_i = (X, Y, \text{carnumber}, \text{description}, \text{starttime}, \text{endtime}) \), where \( X \) and \( Y \) are the longitude and latitude of the accident location; \( \text{carnumber} \) is the number of vehicles involved in the accident; \( \text{description} \) is the description of the accident, mainly including the safety and responsibility of the persons involved, and the damage of vehicles; the \( \text{starttime} \) is the occurrence time of accident; the \( \text{endtime} \) is the time when the road is clear; The duration of the accident is the end time minus the start time. Some samples are shown in Table 1.

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( \text{carnumber} )</th>
<th>( \text{description} )</th>
<th>( \text{starttime} )</th>
<th>( \text{endtime} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>118.110169</td>
<td>24.48329</td>
<td>2</td>
<td>A has a safety problem and B has no fault behavior, the accident results in a right side damage of car A, a left side damage of car B.</td>
<td>2015/01/01 00:09</td>
<td>2015/01/01 00:10</td>
</tr>
<tr>
<td>118.101324</td>
<td>24.50446</td>
<td>2</td>
<td>A and B have a safety problem, the accident results in a left side damage of car A, a right side damage of car B.</td>
<td>2015/01/01 00:09</td>
<td>2015/01/01 00:10</td>
</tr>
</tbody>
</table>
The architecture of our solution

Figure 1 shows the architecture of our proposed solution to accident duration prediction. It contains three main steps: (1) data preprocessing and analysis. Data preprocessing includes data cleaning and transformation. We deal with missing value, format cleaning, and unreasonable value removal. And then we convert the attribute data into the format which can be used by subsequent algorithms. After completing data preprocessing, we analyze the distribution of the traffic accident durations and the relationships between the feature attributes and the duration. (2) Duration classification based on cost-sensitive Bayesian network. According to the data analysis results, some attributes are selected to build the Bayesian network model, the attributes are enough to qualitatively determine whether the duration time is more or less than 30 min. (3) Duration value prediction based on KNN regression. After getting the class label of the accident, KNN regression model is used to get the duration value in each class. In this case, the nearest neighbor selection is performed with the remaining characteristic attributes, and the average value of the accident duration of K neighbors is output as the predicted value.

Data preprocessing

Data quality is related to the performance of the model and the final result closely, so it is necessary for us to clean the data. We deal with the data from three perspectives: first, determine the range of missing values and calculate the ratio of missing values for each field; second, clean the contents which do not match the format; third, remove the outliers, that is, the values which are not accordant with the common sense.

In order to prepare the original data as the input of subsequent algorithms, we need to convert the three attributes “start_time,” “car_number,” “description” according to the following rules:

1. “start_time” is the time when the accident occurs, and it is consist of year, month, day, hour, and minute. We extract three kinds of information from the “start_time”: first, minutes that count from 0 o’clock in that day, which is named as “time”; second, “weekend,” that is, whether the day is on weekend or not, where 1 represents weekend and 0 represents workday; third, the weather of the day, since we can supplement the weather information from online weather service according to the date.

2. “car_number” is the number of cars that are involved in the accident. According to the statistics, there are 3125 accidents of single-car collision and 22,806 accidents of two cars, the number of multicos collision accident is relatively small, especially speaking, the number of accidents that involve three, four and five cars are 1626, 176, and 35, respectively. In order to prevent the learning model from overfitting, we classify the multicos crash into one class. Since most of the algorithm count from 0, the original semantic of the accident is preserved. The attribute ‘car_number’ is mapped as follows: 0 stands for single-car accident, 1 stands for double-cars accident, 2 stands for multicos accident (the number of cars that involved in one accident >2).
3. “description” describes the information of accident. We extract the keywords which have obvious impact on the increase of duration of an accident, and they are: “rear-end,” “rollover,” “rescue,” “tire damage,” “injured.” We then derive “severity” from the description as an ordinal property. According to the analysis of statistical data, the average duration of the “rear-end” accident is lower than that of all accidents, and this kind of accidents is marked as “0.” The average duration of the accidents that are concerned with the keywords “rollover,” “rescue,” “tire damage,” and “injury” is higher than that of all accidents, and these kinds of accidents are marked as “2.” The average duration of the accidents that do not contain the five keywords is the same as that of all accidents, and this kind of accidents is marked as “1.”

Table 2 shows the variable name, data type, converting rule, and the statistics of each attribute:

**Analysis of data**

**Distribution of duration.** We first make a statistical analysis on the duration of accidents. We count how many times when the duration is $x$ minutes while $x$ ranges between [7,120]. The result is shown as Figure 2, where the horizontal ordinate represents the duration minutes of the accident and the vertical ordinate represents the number of accidents. We can see that the distribution of durations is imbalanced. The number of accidents increases rapidly when the duration is between 0 and 20 min; it reaches the highest point when the duration is about 20 min; and then it declines to zero gradually when the duration is larger than 20 min. The number of accidents when the duration is between 15 and 30 min accounts for about 40% of the total accidents. Therefore, in order to classify the duration of accidents effectively, there are two ways to determine the intervals: (1) make a further segmentation for the time interval [15 min, 30 min], so that a balance on amounts for each interval can be kept. (2) Enlarge the time interval [15 min, 30 min], until the number of accidents in the enlarged interval is relatively balanced with that of other intervals. However, an uneven division of the time interval may lead to decrease on classification accuracy since there are more categories and less samples in each category. Therefore, we tend to employ the second solution.

**Analysis of attributes.** We then analyze the four discrete attributes, that is, “weekend,” “severity,” “car_number,” and “weather,” by one-factor variance analysis, in order to find how the change of values on each attribute affects the target “duration.” The result of one-factor variance analysis is shown in Table 3.

We set the significance level to 0.1, and according to Table 3, the values of “Prob > F” on “weekend,” “severity,” and “car_number” are less than 0.1, that is, the significance level, therefore, we have more than 90% confidence that the three features have significant effects on the target duration and the feature “weather” has little significant effect on the target.

In order to further find the relationship between the features, we adopt two-factor analysis of variance to analyze whether the interaction between features will affect the target duration remarkably or not. The result of the two-factor analysis of variance is shown in Table 4.

We set the significance level to 0.1, and according to Table 4, the values of “Prob > F” on “severity*car_number” and “severity*weekend” are less than 0.1,
that is, the significance level, therefore, we have more than 90% confidence that the two combinations have significant effects on the target duration and other combinations have little significant effect on the target.

We then analyze the feature “time” and the target “duration” visually, and the result is shown in Figure 3, in which the x-axis represents the time of the day, y-axis represents the duration. It can be seen from Figure 3 that there is no linear correlation between the time and the duration, but there is no significant changes within a certain time interval, since the traffic flow as well as the traffic management are similar within a certain time, therefore, we can utilize the data with similarity on time for prediction.

**Accident duration classification based on Bayesian network**

By analyzing the distribution of duration values, we can divide the continuous values into several intervals as the target categories reasonably. By analyzing the correlation between the features and the target duration, we can select the useful features, which have significant impact on the prediction of duration. Therefore, in our solution to predicting the duration of new accidents, we classify the duration of accidents into two categories, i.e. more or less than 30 minutes, according to its severity, involved car number and the condition whether it happened in workdays or weekends first.

The reason why we choose 30 minutes as the division point and divide the duration of accidents into two categories is, according to Figure 2 and the analysis in Distribution of duration section, we can see that the distribution of duration is extremely imbalanced, and there are about 40% of samples whose duration is between 15 and 30 min. We have analyzed the drawback if we divide the time interval unevenly. But if we divide the duration evenly with a smaller interval, the problem of imbalanced samples in each category will become especially obvious, the classifier cannot learn the features of the category with little sample sufficiently and the possibility that a test data is classified to the category with little samples is also little. Therefore, we divide the duration of accidents into two categories, that is, less and larger than 30 min, so as to keep the balance of samples in both categories, as well as guarantee that there are enough samples in both categories. In addition, we compare the two ways of interval division in the experiment part, and verify that we model the accidents duration classification as binary classification performs better than multiclassification.

According to the result of two-factor analysis of variance, the interaction between attributes has a significant effect on prediction target. Therefore, we employ the Bayesian network for classification in the first step. The Bayesian network structure is a directed acyclic graph in which nodes represent domain variables and arcs between nodes represent probabilistic dependencies (Koppelman et al., 1994). We will train the Bayesian network by MMPC and K2 algorithms first and then improve the model by adding a cost-sensitive function.

**Training of Bayesian network.** In order to train the Bayesian network, we need to determine the hierarchical order between the nodes in the topology. MMPC is a local discovery and learning algorithm which returns the set of parent nodes of the target variable T, given the target variable T and the dataset D. The MMPC algorithm mainly calls two functions, \text{Ind}(X; T|Z) and \text{Assoc}(X; T|Z). \text{Ind}(X; T|Z) tests the conditional independence of two nodes X and T. \text{Assoc}(X; T|Z) evaluates the correlation strength of node X and T when the evidence Z is given (Koppelman et al., 1994).

After determining the order of nodes by MMPC, we then construct the topology of the Bayesian network by K2 algorithm. The main idea of K2 is described as follows: for each node i in the given

**Table 4. The result of the two-factor variance analysis.**

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum sq.</th>
<th>df</th>
<th>Mean sq.</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>severity × car number</td>
<td>72,376.9</td>
<td>4</td>
<td>18,094.2</td>
<td>56.82</td>
<td>0.0</td>
</tr>
<tr>
<td>severity × weekend</td>
<td>10,973.5</td>
<td>2</td>
<td>5486.8</td>
<td>17.23</td>
<td>0.0</td>
</tr>
<tr>
<td>severity × weather</td>
<td>1023.8</td>
<td>2</td>
<td>511.9</td>
<td>1.49</td>
<td>0.22</td>
</tr>
<tr>
<td>car number × weekend</td>
<td>669.2</td>
<td>2</td>
<td>334.6</td>
<td>0.98</td>
<td>0.38</td>
</tr>
<tr>
<td>car number × weather</td>
<td>1203.9</td>
<td>2</td>
<td>602.0</td>
<td>1.76</td>
<td>0.17</td>
</tr>
<tr>
<td>weekend × weather</td>
<td>193.2</td>
<td>1</td>
<td>193.2</td>
<td>0.56</td>
<td>0.45</td>
</tr>
</tbody>
</table>
order, the $i - 1$ nodes that are in front of node $i$ will be the candidate parent node of $i$, and the Bayesian score of the network structure will be calculated after each candidate parent node is added. The Bayesian score is defined as formula (1), where $pa(x_i)$ denotes the candidate parent node of $x_i$, $q_i$ denotes the number of states of the parent of $x_i$, $r_i$ denotes the number of states of $x_i$, $N_{ij}$ denotes the number of samples when the parent of $x_i$ is in state $j$, $N_{ijk}$ denotes the number of samples when the state of $x_i$ is $k$ and the state of its parent is $j$. If the Bayesian score become higher after adding a candidate parent, the node will be added to the parent set of $i$, and a directed edge from the parent to node $i$ will be added. The loop will continue until the score no longer increases or the maximum number of parents have reached.

$$g(x_i, pa(x_i)) = \frac{n!}{\prod_{j=1}^{q_i} (r_j - 1)! \prod_{k=1}^{r_i} N_{ijk}!}$$  \hspace{1cm} (1)$$

The $T_{type}$ represents the type of the duration, which is discretized into several time intervals. In our solution, we just divide the values of duration into two types $C_0$ and $C_1$, where $C_0$ represents the values that are less than 30 min, and $C_1$ represents the values that are more than 30 min. The topology of the constructed Bayesian network is shown in Figure 4. We can see the interaction relations between the four properties from Figure 4. The property “car_number” has a direct effect on “severity” and “T_type,” and “T_type” are affected by both “severity” and “weekend.”

**Cost-sensitive Bayesian network.** There is a cost-sensitive problem in the prediction of traffic accidents duration in practical application. For example, if a long-lasting accident is estimated incorrectly, vehicles will be misled to select a wrong route, thereby increasing the congestion and road pressure; While if a short-lasting accident is estimated incorrectly, vehicles will plan other routes, only slightly increasing the car energy consumption. So we would like to introduce a cost-sensitive function into the Bayesian network model.

Table 5 shows the cost matrix, in which $C_0$ and $C_1$ have been introduced above; the values of $F(C_0, C_0)$ and $F(C_1, C_1)$ are 0; $F(C_0, C_1)$ represents the cost of misjudging $C_0$ as $C_1$, $F(C_1, C_0)$ represents the cost of misjudging $C_1$ as $C_0$. The value of $F(C_0, C_1)$ and $F(C_1, C_0)$ are calculated as formula (2) shows:

$$F(c_i, c_j | i \neq j) = \begin{cases} \frac{x_j}{x_i}^\beta, & x_i > x_j; \\ \frac{x_i}{x_j}^\alpha, & x_i < x_j; \\ 1, & x_i = x_j \end{cases}$$  \hspace{1cm} (2)$$

Table 5. The cost matrix.

<table>
<thead>
<tr>
<th>Type</th>
<th>$C_0$</th>
<th>$C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_0$</td>
<td>$F(C_0, C_0)$</td>
<td>$F(C_0, C_1)$</td>
</tr>
<tr>
<td>$C_1$</td>
<td>$F(C_1, C_0)$</td>
<td>$F(C_1, C_1)$</td>
</tr>
</tbody>
</table>

In each category, the ratio of samples in category $C_i$ and $x_i$ denotes the ratio of samples in category $C_j$. According to our analysis, we will adjust the value of $\alpha$ and $\beta$ in the cost-sensitive function to make sure that the punishment of $F(C_1, C_0)$ is greater than that of $F(C_0, C_1)$.

Then the cost-sensitive Bayesian network can be achieved by replacing the comparison of $P(C_0|weekend, severity, car_number)$ and $P(C_1|weekend, severity, car_number)$ by that of $R(C_0|weekend, severity, car_number)$ and $R(C_1|weekend, severity, car_number)$, where they can be calculated as follow:

$$R(C_0|weekend, severity, car_number) = P(C_0|weekend, severity, car_number) \ast F(C_0, C_0) + P(C_1|weekend, severity, car_number) \ast F(C_1, C_0)$$

$$R(C_1|weekend, severity, car_number) = P(C_0|weekend, severity, car_number) \ast F(C_0, C_1) + P(C_1|weekend, severity, car_number) \ast F(C_1, C_1)$$

**Predicting the value of accident duration based on KNN**

After the duration of a new accident is classified into a determined range, we then predict the detailed value...
of the duration based on KNN. In the selection of K nearest accident samples A and B, the distance between two accident samples A and B is defined as formula (5) shows:

\[
\text{distance} (A, B) = (1 - \lambda) \times \text{timeDist}(A, B) + \lambda \times \text{spaceDist}(A, B),
\]

where \( \text{timeDist}(A, B) \) and \( \text{spaceDist}(A, B) \) are defined as formula (6) and (7) shows:

\[
\text{timeDist}(A, B) = |T_A - T_B|,
\]

\[
\text{spaceDist}(A, B) = \frac{R \times \cos^{-1}(\sin(lat_A) \times \sin(lat_B) \times \cos(lon_A - lon_B) + \cos(lat_A) \times \cos(lat_B)) \times \pi}{180},
\]

where \( T_A \) and \( T_B \) are the starting time of accident A and B, respectively, where \( R \) denotes the radius of the earth, \( \text{lon}A, \text{lat}A \) and \( \text{lon}B, \text{lat}B \) are the coordinates of A and B, respectively.

After K neighbors are selected according to formula (5), the average value of K neighbors are generated as the predicted value.

### Experiment

After data cleaning and processing, there are 37,712 traffic accident samples from Xiamen City in 2015. We use 27,712 samples as the training set, and the remaining 10,000 samples as the test set. In each experiment, we use the whole training set to train the model, and we randomly select 1000 samples from the test set as six subdatasets and record the performance. The metric that we use for evaluating the performance of the classification models is Accuracy rate, for the predicting model is MAPE, and their definitions are given as formula (8) and (9).

\[
\text{Accuracy rate} = \frac{n}{N},
\]

where:

\( N \) denotes the number of samples in test dataset; \( n \) represents the number of samples that are classified correctly in test dataset;

\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\text{observed}_i - \text{predicted}_i}{\text{observed}_i} \right|,
\]

where:

\( N \) denotes the number of samples in test dataset; \( \text{observed}_i \) denotes the real value of the accident duration; \( \text{predicted}_i \) denotes the predicted value of the accident duration.

In the following, we will first determine the parameters involved in our approach, and then compare our models with other relative ones.

### Determining the parameters

First, we aim to discretize the duration values into several time intervals reasonably to avoid the imbalance of samples. According to the analysis on distribution of duration values, we have two solutions: (1) divide the range of duration every 15 min, therefore, we can map the duration values into five categories, that is, \([0,15], [15,30], [30,45], [45,60], [60,\infty]\); (2) divide the range of duration into two categories, that is \([0,30]\) and \([30,\infty]\).

In order to decide which solution would be better, we introduce a base classifier, and compare it with our classification model, that is, cost-sensitive Bayesian network. The base classifier predicts the category that a test data belongs to according to the highest proportion of categories in the training dataset.

Figure 5(a) shows comparison of two classification models with interval of 15 min, while Figure 5(b) shows that with interval of 30 min, in both of which x-axis denotes the six randomly generated subdatasets in the test set, the y-axis denotes the accuracy rate of classification. According to the experimental results, we can see that the performance of two classification models are almost the same for each test set in Figure 5(a) while the cost-sensitive Bayesian network outperforms much than base classifier in Figure 5(b), therefore, it is reasonable to classify the duration values into two categories, that is \([0,30]\) and \([30,\infty]\), since it can deal with the imbalance of samples in each categories effectively.

Next, we aim to decide the parameters \( \alpha \) and \( \beta \) in the cost-sensitive function. According to formula (2), when \( \alpha = \beta, F(C_1, C_0) = F(C_0, C_1) \). But in our scenario, in order to make the model biased towards the accidents with duration value more than 30 min, the value of \( F(C_0, C_1) \) should be less than that of \( F(C_1, C_0) \), so we need to adjust the \( D \)
value between $\alpha$ and $\beta$ as much as possible. In the experiments, we reduce the value of $\alpha$ from 1.0, and increase the value of $\beta$ from 1.0, and we observe the changes on the accuracy rate of classification on both types of accidents, as well as the misclassification rate on the second type of accidents.

Figure 6 shows the experimental result of determining $\alpha$ and $\beta$. We can see that, the total accuracy rate increases slowly first, reaches the highest point when $\alpha = 0.5$ and $\beta = 1.5$, and then it gradually decreases. When the total accuracy rate is the highest, the misclassification rate of the accidents with long duration is 12.65%. And when $\alpha = 0.5$, $\beta = 2.2$, though the total accuracy rate is 74.6%, the misclassification rate of the accidents with long duration is only 1.19%, thus we set $\alpha$ and $\beta$ values to 0.5 and 2.2, respectively.

Next, we will determine the parameter $\lambda$ in formula (5). We use MAPE to evaluate the performance of duration value prediction based on KNN. Figure 7 shows the result, in which $x$-axis is the value of $\lambda$, $y$-axis is the MAPE result, the lines in different colors represent the performance under different values of $k$. We can see from Figure 7 that, when $\lambda$ is in $[0, 0.1]$, the change of MAPE presents a downward trend; when $\lambda$ is in $[0.1,0.3]$, the values of MAPE are relatively stable with some small fluctuations; when $\lambda$ is in $[0.3,0.5]$, MAPE tends to increase. Therefore, when $\lambda$ is in $[0.1, 0.3]$, the prediction model presents a better performance, and finally we set $\lambda$ to 0.3.

Finally, we aim to determine the value of $k$ in the KNN model. Figure 8 shows the MAPE under different values of $k$, while the lines in different colors
represents the six test datasets. We can see from Figure 8, generally speaking, for all the test sets, when $k$ is increased from 1 to 10, the value of MAPE reduces quickly, and when $k$ is greater than 10, the value of MAPE increases gradually. So we set $k$ to 10 to get a better performance.

**Verifying the models**

**Traditional Bayesian network vs. cost-sensitive Bayesian network**

First, we aim to compare the traditional Bayesian network with our improved cost-sensitive Bayesian network, from the perspective of total accuracy rate and the misclassification rate of the accidents with long duration. The results are shown in Table 6, we can see that the accuracy rate increases 3.8%, while the misclassification rate of accidents with long duration decreases 13.83% if we use cost-sensitive Bayesian network.

In order to find the influence of cost-sensitive function on the final prediction, we measure the MAPE without and with cost-sensitive function, in which the time range is divided further into 0–15, 15–30, 30–45, 45–60, and 60–$\infty$. The experiments run on six test datasets. The result is shown in Table 7. It shows that we have a little loss on prediction accuracy if we use cost-sensitive function, since it tends to have a larger prediction value for short duration of accidents, but it can greatly decrease the possibility of misclassifying a large duration to the first category, that is, [0–30]. But generally speaking, the loss can be acceptable, according to Table 7, MAPE only increases 0.0265 on average for accidents with less than 30 min duration, while decreases 0.076 on average for accidents with more than 30 min duration.

**Comparison of classification models**

Since we find the interaction effects between the input features, we choose Bayesian network as the classification model. In this section, we aim to compare our cost-sensitive Bayesian network with three other classification models, which are Naive Bayesian, decision tree and random forest. Experiments run on six test datasets, and accuracy rate is used for evaluating the performance of models, and the result is shown in Figure 9.

### Table 6. Comparison between Bayesian network and cost-sensitive Bayesian network.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy rate</th>
<th>Misclassification rate of accident with long duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian network</td>
<td>71.7%</td>
<td>26.48%</td>
</tr>
<tr>
<td>Cost-sensitive</td>
<td>75.5%</td>
<td>12.65%</td>
</tr>
</tbody>
</table>

### Table 7. MAPE of the Bayesian network without and with cost-sensitive function.

<table>
<thead>
<tr>
<th>Test data</th>
<th>0–15</th>
<th>15–30</th>
<th>30–45</th>
<th>45–60</th>
<th>&gt;60</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE without cost-sensitive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DataSet1</td>
<td>0.518</td>
<td>0.314</td>
<td>0.182</td>
<td>0.355</td>
<td>0.572</td>
</tr>
<tr>
<td>DataSet2</td>
<td>0.520</td>
<td>0.315</td>
<td>0.177</td>
<td>0.343</td>
<td>0.573</td>
</tr>
<tr>
<td>DataSet3</td>
<td>0.540</td>
<td>0.336</td>
<td>0.230</td>
<td>0.386</td>
<td>0.582</td>
</tr>
<tr>
<td>DataSet4</td>
<td>0.510</td>
<td>0.330</td>
<td>0.212</td>
<td>0.341</td>
<td>0.557</td>
</tr>
<tr>
<td>DataSet5</td>
<td>0.581</td>
<td>0.332</td>
<td>0.209</td>
<td>0.327</td>
<td>0.571</td>
</tr>
<tr>
<td>DataSet6</td>
<td>0.587</td>
<td>0.326</td>
<td>0.215</td>
<td>0.339</td>
<td>0.564</td>
</tr>
<tr>
<td>Average</td>
<td>0.532</td>
<td>0.326</td>
<td>0.204</td>
<td>0.349</td>
<td>0.570</td>
</tr>
<tr>
<td>MAPE with cost-sensitive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DataSet1</td>
<td>0.540</td>
<td>0.340</td>
<td>0.167</td>
<td>0.244</td>
<td>0.514</td>
</tr>
<tr>
<td>DataSet2</td>
<td>0.558</td>
<td>0.349</td>
<td>0.145</td>
<td>0.240</td>
<td>0.493</td>
</tr>
<tr>
<td>DataSet3</td>
<td>0.551</td>
<td>0.340</td>
<td>0.217</td>
<td>0.241</td>
<td>0.496</td>
</tr>
<tr>
<td>DataSet4</td>
<td>0.558</td>
<td>0.348</td>
<td>0.161</td>
<td>0.232</td>
<td>0.500</td>
</tr>
<tr>
<td>DataSet5</td>
<td>0.600</td>
<td>0.342</td>
<td>0.136</td>
<td>0.224</td>
<td>0.493</td>
</tr>
<tr>
<td>DataSet6</td>
<td>0.603</td>
<td>0.339</td>
<td>0.141</td>
<td>0.223</td>
<td>0.495</td>
</tr>
<tr>
<td>Average</td>
<td>0.566</td>
<td>0.343</td>
<td>0.161</td>
<td>0.234</td>
<td>0.499</td>
</tr>
</tbody>
</table>

Figure 7. MAPE under different values of $\lambda$ and $K$.

Figure 8. MAPE under different values of $K$ in six test datasets.

Table 6. Comparison between Bayesian network and cost-sensitive Bayesian network.

Table 7. MAPE of the Bayesian network without and with cost-sensitive function.
We can see that, cost-sensitive Bayesian network performs the best in all the test datasets, and in the first three test datasets, naive Bayes performs the second, while in the latter three test datasets, random forest performs the second. Therefore, we can make a conclusion that it is reasonable to choose cost-sensitive Bayesian network for the classification of duration values.

Comparison of regression models

In this part, we aim to compare our KNN model with linear model for predicting duration value. Experiments run on six test datasets and MAPE is used for evaluation. The result is shown in Figure 10. The MAPE of KNN model is 0.391 in the third dataset, which is the lowest in the six test datasets, and the MAPE of linear model is 0.577 under the same condition. The MAPE of KNN model is 0.471 in the first dataset, which is the highest in all datasets, while the MAPE of linear model is 0.670 under the same conditions. We can see that KNN always performs better than linear model, therefore it is reasonable to choose KNN model as the regression model to predict the duration of traffic accidents.

Conclusions

In this article, we propose a Bayesian network-weighted KNN model to predict the duration of accidents, using the traffic accident data of Xiamen, China, from January to December 2015, and experimental results show that the algorithm can improve the prediction accuracy in cost-sensitive way. In the future, we aim to explore visualization techniques.

Figure 9. Comparison on accuracy rate between cost-sensitive Bayesian network and other three classical classification models.

Figure 10. Comparison on MAPE between weighted KNN and linear regression model.
(Liao et al., 2018) for data analysis, and investigate on how to use sensors to take more valuable information on the intelligent traffic management system.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

The research is supported by National Natural Science Foundation of China (no. 61772560, 61876190, and 61872306) and Scientific Research Project for Professors in Central South University, China (no. 904010001).

**References**


