IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 21, 2019, accepted September 12, 2019, date of publication September 27, 2019, date of current version October 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942647

A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China

CHENG ZHANG, YUE SHU, AND LIXIN YAN
School of Transportation and Logistics, East China Jiaotong University, Nanchang 330013, China

Corresponding author: Lixin Yan (yanlixinits@126.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 51805169.

ABSTRACT With the rapid development of the social economy and accelerating urbanization, the total number of motor vehicles continues to grow at a high rate. Roads in large- and medium-sized cities are becoming increasingly congested, which leads to frequent traffic accidents. To enhance road traffic safety and reduce the traffic accident rate, effectively identifying accident black spots is of great importance. In this study, the data from traffic accidents on the Lianfeng Middle Road, Yinzhou District, Ningbo City were selected for the analytical dataset, and eight impact factors (holiday, day of week, time, rush hour traffic, accident location type, accident type, weather, responsibility and black spot) were set. The improved K-means clustering algorithm was proposed to solve the shortcomings of the traditional algorithm, which is susceptible to outliers and initial clustering centres. Through this algorithm, the traffic accidents in the dataset were divided into two categories: black spots and non-black spots. Then, using the updated dataset, we employed a Bayesian network to construct a black spot identification model, and applied other widely used algorithms (the ID3 decision tree, logistic regression and support vector machine) for comparison. The values of the ROC area, TP rate, FP rate, precision, recall, F-measure and accuracy reached 0.618, 0.668, 0.580, 0.650, 0.668, 0.590 and 0.668, respectively, which showed that the Bayesian network was the best model to effectively identify road accident black spots. Moreover, a bivariate correlation model was applied to verify the correlation between the impact factors and black spots. The results indicated that the accident location type, accident type, time, and responsibility had significant correlations with black spots, which had a value of sig<0.05. The conclusions could provide reference evidence for the identification and prevention of traffic accident black spots to significantly contribute to traffic safety.

INDEX TERMS Traffic safety, K-means clustering, Bayesian networks, black spot identification.

I. INTRODUCTION

The global status reports on road safety released by the World Health Organization in December 2018 emphasized that the number of road traffic deaths per year has reached 1.35 million [1]. Traffic injuries are becoming the leading killer of people aged 5 to 29 years old. The data show that low- and middle-income countries bear the greatest burden of road traffic fatalities and injuries. According to the Chinese annual statistics report on road traffic accidents, compared to 2016, the total number of traffic accidents in 2017 fell by 4.6%; however, the death and injury rates remained high, which were 4.59 per 100,000 people and 15.08 per 100,000 people, respectively [2]. Taking Yinzhou, Ningbo as an example, with the continuous high-speed growth of motor vehicles, traffic accidents in this district occur frequently, which is detrimental for economic development and urbanization. To improve road safety, it is of great importance to use traffic accident data resources to develop targeted traffic safety measures that will effectively reduce the accident rate. Taking into account the time and money problems, it is impractical to improve road safety at all collision points, so black spot identification technology is widely studied to identify the accident-prone locations. Additionally, it is an economical and efficient way to decrease traffic accidents by analysing the cause of black spots.

Road traffic accident black spots, also known as dangerous road spots, refer to the road spaces where the incidence and

The associate editor coordinating the review of this manuscript and approving it for publication was Yue Cao.

consequences of traffic accidents are significant over a period of time. Thus far, there is no unified meaning of the road accident black spot that is generally accepted worldwide. Several classic definitions are as follows. The UK defined the black spot as the location where four accidents occur in a year on a 100-meter road. In Norway, the black spot refers to a 100-meter road which has traffic events involving more than four casualties in the past four years. Previous studies also have different opinions on the definition of the black spot recognition targets. Similar types of traffic accidents tend to occur frequently somewhere on a road. Shen *et al.* [3] called these black spots ‘‘accident black spots’’. Meuleners *et al.* [4] regarded a location with one or more accidents as a black spot. Murat and Cakici [5] identified a black spot as a dense spot around a cluster centre. In this study, the definition of black spots will be based on the official definition from where the accident data were obtained.

In the identification process of black spots, three main methods can be used: screening, clustering and crash prediction methods [6]. Black spots are easily screened according to the different threshold values for the safety status of roads. However, in practice, it is easy to overlook some triggering factors (road conditions, accident severity, etc.), so this method of black spot recognition cannot provide suggestions for the projects needed to reduce the safety hazards. Therefore, in recent years, machine learning algorithms have been widely used in road accident prediction. They can effectively classify datasets and establish a link between the factors and the severity of the traffic events.

Machine learning algorithms include classification, regression, clustering, and dimensionality reduction. The K-means algorithm is a dynamic clustering algorithm based on partitioning. As a commonly used data-identification approach, there is no particular limitation on the clustering range, and as long as multiple datasets are independent of each other, they can be run in parallel. Gupta *et al.* [7] proved that it is effective to implement black spot detection in the lane using the K-means algorithm. Zhong and Liu [8] proposed a K-means spatial clustering algorithm to achieve the unification of spatial similarity and entity attribute similarity. Other clustering algorithms have also been used to identify road accident black spots. For example, Murat and Cakici [5] used fuzzy clustering to identify accident black spots and incorporated entropy to evaluate the features.

The Bayesian network (BN) classification algorithm is flexible in classifying accident datasets. It utilizes the previous data information and adds the experience of the decision makers. Moreover, BNs can be used to study the relationships among variables to make predictions without any presuppositions. Deublein *et al.* [9] analysed the case of the Austrian Highway Research Network to demonstrate the usefulness of the Bayesian model for predicting the severity of injuries. Mujalli *et al.* [10] used BNs to improve the classification of unbalanced accident datasets. Mbakwe *et al.* [11] combined Delphi technology with BNs to predict road traffic accidents in developing countries. Other classification algorithms, such as the ID3 decision tree (ID3), logistic regression (logistic) and support vector machine (SVM), have been employed to classify different types of datasets in previous studies.

Considering the uncontrollability and uncertainty of black spots and trying to approach the true definition of black spots, scholars have tended to combine multiple methods to enhance the identification accuracy in recent years. Ali and Tayfour [12] used artificial neural networks and regression techniques to predict traffic accident casualties. The comparison of the predictions with recorded data was very favourable. Lu [13] employed a quality control method based on hierarchical storage management and the empirical Bayesian model to determine the threshold of black spot recognition. Dereli and Erdogan [14] applied model-based spatial statistics to determine traffic accident black spots. By comparing the negative binomial regression, Poisson regression and empirical Bayesian models used in the study, it was found that the empirical Bayesian approach provides the best results in terms of consistency and accuracy. Xiao [15] proposed an SVM and K-nearest neighbour ensemble learning method to improve the robustness of traffic incident detection. Debrabant *et al.* [16] used the optimized kernel density clustering method and the buffer method to determine black spots for accident occurrence points, road sections, and regions.

In the current study, K-means clustering has been applied to determine the road traffic accident black spots, and BNs have been mostly used to classify and predict traffic accidents. Few studies have combined the two methods to study the classification and prediction of black spots. Therefore, in this work, a novel identification model combining K-means with BNs was proposed. Three contrast models were established to prove that the methodology is effective and optimal.

This study is organized into the following steps. First, the traffic accident data collection and processing, including the data screening and variable setting, are introduced. Then, an improved K-means clustering algorithm combined with BNs, ID3, logistic and SVM methods are selected to establish the accident black spot identification models. Third, the accuracies of four models are evaluated by using a receiver operating characteristic curve (ROC) and other evaluation indicators. Finally, a correlation analysis model is introduced to find the key factors that have a strong correlation with traffic accident black spots.

II. METHODOLOGY

A. K-MEANS CLUSTERING ALGORITHM

MacQueen [17] first proposed the K-means algorithm. Because of its simplicity and efficiency, it has been widely applied and studied in recent years. The specific process of the K-means algorithm is as follows.

The traditional K-means algorithm has some drawbacks: The number of initial clusters (K) needs to be specified in advance. However, in practice, the value of K is hard

Algorithm 1 Traditional K-Means Algorithm

```
Input: Dataset $D = \{x_1, x_2, \ldots, x_m\}$; Cluster number, $K$.
Output: Cluster division $C = \{C_1, C_2, \ldots, C_K\}$
Select K samples from D as the initial mean vector randomly $\{\mu_1, \mu_2, \ldots, \mu_K\}$
Initialization: $C_i \leftarrow \emptyset (1 \leq i \leq k)$
Repeat
  For $j = 1, 2, \ldots, m$ do
    $d_{ji} = \|x_j - \mu_i\|_2$ /* Calculate the distance between the sample $x_j$ and each mean vector $\mu_i (1 \leq i \leq k)$*/
    $\lambda_j = \arg\min_{i \in \{1,2,\cdots,k\}} d_{ji}$ /* Determine the cluster tag of $x_j$ based on the closest mean vector */
    $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$ /* Divide the sample $x_j$ into the corresponding cluster */
  End for
  For $i = 1, 2, \ldots, K$ do
    $\mu_i' = \frac{1}{C_i} \sum_{x \in C_i} x$ /* Calculate the new mean vector */
    If $\mu_i' \neq \mu_i$ then
      Update the current mean vector $\mu_i$ to $\mu_i'$
    Else
      Keep the current mean vector unchanged
    End if
  End for
Until the current mean vector is not updated
```

to determine. Moreover, the result is very sensitive to the initial clustering centre. Because the selection of the initial clustering centres is random, different centres will eventually lead to different clustering results. It is easy to fall into a local minimum solution. Concurrently, the K-means algorithm is susceptible to isolated points (also known as noise data), which leads to complexity in its application to large datasets.

To speed up the clustering of data and improve the accuracy of black spot recognition, an improved K-means algorithm is described as following: 1) remove the isolated traffic accident points based on distance; 2) initialize the coordinates of the cluster centre; 3) call the traditional K-means algorithm. The advantages of the improved algorithm include eliminating noise data, which affect the accuracy of recognition, and not needing to specify the value of K in advance. These improvements can enhance the efficiency and accuracy of clustering.

B. BAYESIAN NETWORKS

Bayesian network was first proposed by Judea Pearl (1988). It consists of a direct acyclic graph (DAG) and a conditional probability table (CPT). The construction of BNs can be divided into two stages: 1) use the BN learning algorithm to form a network structure composed of all attribute and class variables; 2) use the BN inference algorithm to calculate the probabilities of the class variables when the values of attribute variables are given.

Algorithm 2 Improved K-Means Algorithm

```
Input: Dataset $D = \{x_1, x_2, \ldots, x_m\}$; Screening distance, $r$; Threshold, $N_0$; Mutation distance, $M \gg m$.
Output: Cluster division $C = \{C_1, C_2, \ldots, C_K\}$
For $x_j \in D$ do
  Calculate the number of data points ($N$) with $x_j$ as the centre and $r$ as the radius
  If $N \leq N_0$ then
    Remove $x_j$ from $D$
  Else
    $D' \leftarrow x_j$
  End if
End for
For $x_j \in D'$ do
    $m_{X_j} = |X_j - X_{j-1}|; m_{Y_j} = |Y_j - Y_{j-1}|$ /* Arrange the samples $x_j(X, Y)$ by coordinate size; calculate the coordinate distance between two adjacent data points */
    If $m = M$ then
      $x_j$ and $x_{j-1}$ implement jumps between the clusters
      Mark the number of times $m$ equals $M$ as $k$, and the number of clusters is $K = k + 1$
      $C_i = \left[\frac{X_j + X_j'}{2}, \frac{Y_j + Y_j'}{2}\right], (1 \leq i \leq k)$ /* Calculate the initial cluster centre coordinates based on the two data points at the edge of each cluster */
        $C \leftarrow C_i$
      Call the traditional K-means algorithm
    End if
  End for
```

1) THE BN LEARNING STRUCTURE BASED ON THE TAN CLASSIFIER

The TAN (tree-augmented naïve Bayes) classifier is based on the Naïve Bayes classifier. The structure of a BN based on the TAN classifier is shown in Fig.1.

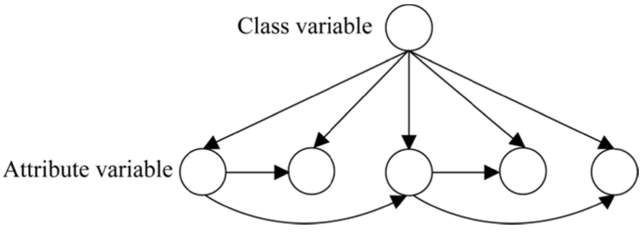


FIGURE 1. The structure of a BN based on the TAN classifier.

Naïve Bayes assumes that attributes are independent and equally important. However, this assumption might not be correct in reality. Mutual information (MI) is a feature selection algorithm used to represent the correlation between two variables. Using the MI value as a weight reflects that different attribute variables have dissimilar effects on the classification and largely eliminates the influence of the independence assumption on the classification effects. The typical TAN is a learning algorithm that constructs a TAN classifier using the conditional mutual information. Hypothesis the training set D is complete. A_i and $A_j (i, j = 1, 2 \ldots n)$ are the attribute variables. C is the class variable. Thus, the learning of the BN structure can be calculated as follows.

Step 1: Discrete variables

Each variable is divided into states. The states are defined as a_r, a_q and c_k, where r, q and k are the number of states of A_i, A_j and C.

Step 2: Calculate the value of *MI*

MI represents the mutual information between attribute variables. $MI(A_i, A_j|C)$ represents the *MI* between a_i and a_j, given C. The value of *MI* can be expressed by Eq.(1).

$$MI(A_i, A_j|C) = \sum_{A_i, A_j, C} P(a_r, a_q, c_k) \log \left[\frac{P(a_r, a_q|c_k)}{P(a_r|c_k)P(a_q|c_k)} \right] \tag{1}$$

Step 3: Establish an undirected graph

Establish a weighted completely undirected graph. The weight of an arc is the value of $MI(A_i, A_j|C)$.

Step 4: Construct the maximum weight tree

The maximum weight tree is constructed according to the principle of not generating a loop. The node pairs are taken out in descending order of the MI value of each attribute pair until all the arcs $(n - 1)$ have been selected.

Step 5: Establish a directed graph

Select a root node and set the outward connection direction, and then convert the undirected tree into a directed tree. Add the classified variable node and its directed connection to each attribute node to construct the network structure based on TAN.

2) CALCULATION OF THE PROBABILITY DISTRIBUTION

$X = \langle \langle X, T \rangle, P \rangle$ represents a BN with X nodes. $\langle X, T \rangle$ is a directed acyclic graph (G). The set of nodes, $X = (X_1, X_2, \ldots, X_n)$, represents the set of variables, and the directed edges (T) between the nodes represents the correlations between the variables. For directed edges (X_i, X_j), X_i is the parent node of X_j, and X_j is the child node of X_i. A node without a parent is called the root node. P denotes the conditional probability distribution associated with each node. According to the conditional independence assumption of the BN, the CPD can be described by $P(X_i|Parent(X_i))$, which expresses the correlation between the node and its parent node. Combining the prior probability of the root nodes with the conditional probability, a joint probability distribution containing all the nodes can be obtained. The inference formula is shown in Eq.(2).

$$\begin{aligned} P(X_1, X_2, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i|X_1, X_2, \ldots, X_{i-1}) \\ &= \prod_{i=1}^{n} P(X_i|Parents(X_i)) \end{aligned} \tag{2}$$

where $i = 1, 2 \ldots n$, and $Parents(X_i)$ represents the parent node set of X_i.

C. CLASSIFICATION ALGORITHMS

The ID3 (iterative dichotomiser 3) algorithm is a decision tree learning algorithm. It takes the decline speed of the information entropy as the criterion for selecting the test attributes. The attribute with the largest information gain at each node is selected as the best classification criterion. Consider the variable,$X = \{x_1, x_2, \ldots, x_n\}$ which corresponds to the probability of the set being $P = \{p_1, p_2, \ldots, p_n\}$. The information entropy is the expected value of information. The information gain is the information entropy ($H(D)$) of the class variable minus the information entropy ($H(X)$) of each attribute variable. The calculation formulas for information entropy and information gain are as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{3}$$

$$Info\_Gain(D, X) = H(D) - H(X) \tag{4}$$

The logistic (logistic regression) algorithm is a regression algorithm that is mostly used to perform linear binary classification. It requires few assumptions and the model results are highly interpretable. The algorithm can be described as in Eq.(4). It classifies the dataset by finding a suitable function and establishing a regression formula.

$$Y(y = 1) = \frac{1}{1 + e^{-W^T X}} \tag{5}$$

where Y represents the dependent variables; X represents the independent variables; and W represents the weight vectors from data learning.

An SVM (support vector machine) is an algorithm for classifying linear and nonlinear data. Its basic idea is to map spatially inseparable data through a kernel function to a high-dimensional space. The SVM exhibits many unique advantages in solving small sample and high dimensional pattern recognition problems.

D. THE EVALUATION INDICATORS

The ROC curve (receiver operating characteristic curve, sensitivity curve): The points on the curve represent the reflection of the same signal stimulus. It is a curve drawn using a series of different two-category methods (the demarcation value or decision threshold). The ordinate is the true positive rate (*sensitivity*). The abscissa is a false positive rate ($1-$*specificity*). An ROC curve analysis is a statistical method that can effectively describe the overall test performance of a classification model.

The AUC (area under the curve): The AUC is the area enclosed by the ROC curve and the coordinate axis. Since the ROC curve is generally above the reference line ($y = x$), the AUC value ranges between 0.5 and 1. As a value, the AUC can more intuitively evaluate the effect of a classifier.

When the AUC value is 1.00, a perfect test is described. When the AUC value is 0.50, it is described as a valueless test.

The descriptions of the other five detailed evaluation indicators are as follows. The true positive rate (*TPR*) measures sensitivity and is defined as the proportion of the samples that are actually positive and are predicted to be positive. The false positive rate (*FPR*) equals 1-specificity and is defined as the proportion of the samples that are actually negative and are predicted to be positive. The larger the ratio is of the TP rate to the FP rate, the better the classification effect of the method is. Precision (*P*) represents the proportion of the samples that are predicted to be positive in the positive case. Recall (*R*) is a measure of coverage and is equal to sensitivity. The F-measure (*F*) can comprehensively evaluate the precision and recall indicators. When the value of F is higher, the classifier is more effective. Accuracy (*ACC*) is a common indicator. It indicates the ratio of the number of correctly classified samples to the total number of samples. In general, a higher value of *ACC* shows a better performance of the classifier. The indicators are obtained by Eqs.(6-11).

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F = \frac{2P \times R}{P + R} \tag{10}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where *TP* is the number of samples that are actually positive and predicted to be positive; *FP* is the number of samples that are actually negative but predicted to be positive; *TN* is the number of samples that are actually negative and predicted to be negative; *FN* is the number of samples that are actually positive but predicted to be negative.

III. DATA COLLECTION AND PROCESSING

A. DATA COLLECTION

The traffic accident alarm data were collected from the big data platform of APP called "Yinzhou Traffic Police" before the study. It contains 12 items: the event number, latitude and longitude coordinates, event time, squadron event, alarm, event position, event type, event cause, weather conditions, event environment and scene photos of the event. Excluding the duplicate data, 37,654 valid events from the fourth quarter of 2016 were retained.

Lianfeng Middle Road is one of the busiest regions in the Yinzhou District. According to Yinzhou traffic accident data reports, the amount of traffic events that occurred on Lianfeng Road and the surrounding areas is approximately 1000 per quarter in recent years, and the event types are complicated. Therefore, Lianfeng Middle Road was taken as the object road. The area was selected according to the longitude and latitude (east longitude: 121.4739° to 121.5053°, north latitude: 29.8654° to 29.8792°). The total number of accidents in the area was 1005. The location map and accident point map of the Lianfeng Middle Road are shown in Figs. 2 and 3.

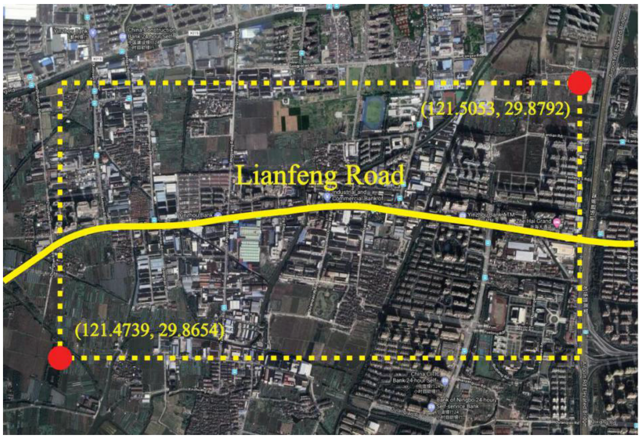


FIGURE 2. The location map of the Lianfeng middle road.

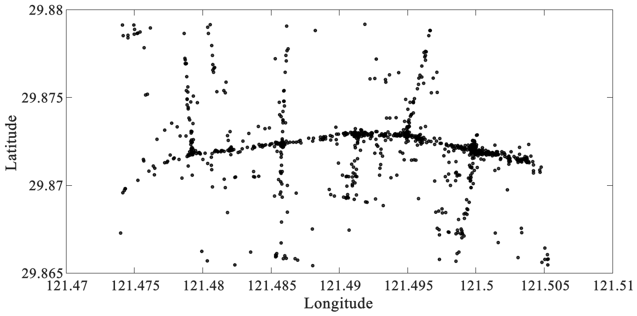


FIGURE 3. The accident points map of the Lianfeng middle road.

B. DATA PROCESSING

The traffic accident data from the fourth quarter of 2016 for the Lianfeng Middle Road area were used in this study. Nine variables were set based on the principle of reducing the model complexity of and improving the accuracy of the model. Eight impact factors (HOL, DAY, TIM, TRH, ACLT, ACT, WEA, and RESP) were defined as attribute variables. HOS was regarded as the class variable.

Considering the accuracy of the analysis results, the data must be processed before use. The traffic accidents were divided into two types (0-hot spot and 1-non-hot spot). The time was divided into day (6:00-18:00) and night (18:00-6:00), and traffic rush hour was defined according to the standards in China. There were three types of accident locations and eight types of accidents. The general information of the selected variables is shown in Table 1.

IV. RESULTS

A. THE RESULTS OF K-MEANS CLUSTERING

The definition of an accident black spot in the Yinzhou traffic police department is as follow. Based on the quarterly

TABLE 1. **Variable descriptions.**

ID	Variable name	Identification	Type	Description
1	Holiday	HOL	Categorical	0, No; 1, Yes
2	Day of week	DAY	Categorical	0, Weekend (Saturday to Sunday); 1, Weekday (Monday to Friday)
3	Time	TIM	Continuous	0, Night (18:00–6:00); 1, Day (6:00–18:00)
4	Traffic rush hour	TRH	Continuous	0, No; 1, Yes (6:30-8:30; 16:30-18:30)

5	Accident location type	ACLT	Categorical	0, Road intersection;
				1, Road section;
				2, residential area or parking lot

6	Accident type	ACT	Categorical	1, Motor vehicles and motor vehicles; 2, Motor vehicles and non-motor vehicles; 3, Motor vehicles and pedestrian; 4, No-motor vehicles and non−motor vehicles; 5, Non-motor vehicles and pedestrian; 6, Single vehicle; 7, Traffic escape; 8, Other
7	Weather	WEA	Categorical	1, Sunny; 2, Rainy; 3, Cloudy; 4, Snowy
8	Responsibility	RESP	Categorical	0, Illegal; 1, Non-illegal; 2, Unclear
9	Black spot	BLS	Categorical	0, No; 1, Yes

average, an area with more than 25 traffic accident points within a radius of 50 metres is considered a potential black spot. First, the improved K-means algorithm was used to identify the black spots on Lianfeng Middle Road. Repeating the experiment several times, the position of the cluster centres did not change, and the distribution of the cluster centres was regular. As shown in Fig.4, 334 black spots were selected from 1005 accidents, and the dataset was divided into six clusters.

In the application of the accident data from Lianfeng Middle Road, the number of classifications (K) is difficult to determine. Therefore, according to the above result of the improved algorithm, when K was specified in advance as six, it was easy to obtain the position of the black spots by the traditional K-means algorithm. However, the distribution of the cluster centres is random and irregular. It could be found that the clustering is inaccurate through repeated experiments. The distribution of the black spots based on the traditional K-means algorithm is shown in Fig.5. Thus, the results show that the improved algorithm could accurately and efficiently identify the cluster centres.

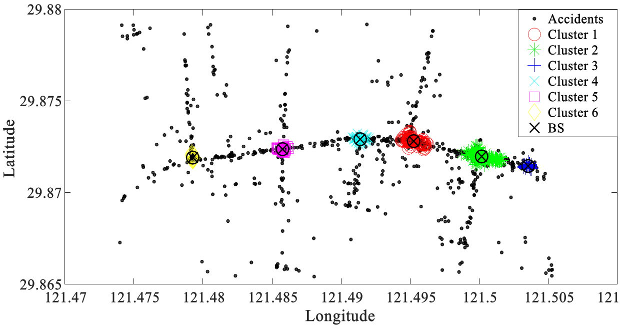


FIGURE 4. **Distribution of the black spots based on the improved K-means algorithm.**

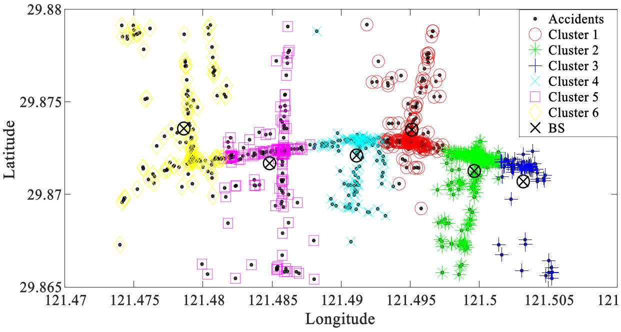


FIGURE 5. **Distribution of the black spots based on the traditional K-means algorithm.**

B. BAYESIAN NETWORK STRUCTURE

The Weka software (Witten and Frank, 2013) was applied to construct the BNs based on the TAN algorithm. From the dataset, 70% of the data were held for training the BNs, and the other 30% were used for testing. The structure of a BN is shown in Fig.6; it contains 9 nodes and 15 directed arcs. The nine nodes match the eight attribute variables and one class variable in Table 1.

C. THE CLASSIFICATION RESULTS

1) THE ROC CURVE

In addition to the BN model above, the ID3, logistic and SVM models were also established for comparison. The ROC curves for the four classification models are shown in Fig.7, where the X-axis represents 1-specificity and the Y-axis represents sensitivity. All the ROC areas exceeded 50%, which suggests that the four models were valid.

As shown in Fig.8, the AUC value for the ID3 and SVM algorithms reached 0.58 and 0.54, respectively, and the best AUC values obtained by the BN and logistic algorithms were

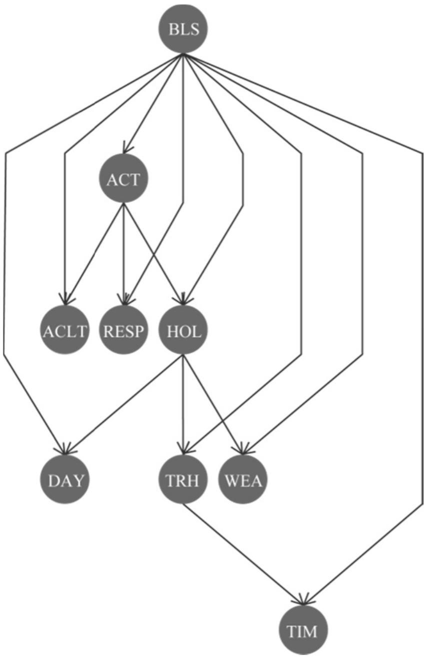


FIGURE 6. The structure of a BN.

both 0.62. De Oña *et al.* [18] built a BN using three different score metrics (BDe, MDL and AIC) to detect the injury severity in traffic accidents. They obtained ROC areas of 62% when using the BDe score, 61% when using the MDL score and 59% when using the AIC score. Thus, the AUC value obtained by the BN in this paper was within the range of the ROC area found by De Oña *et al.*

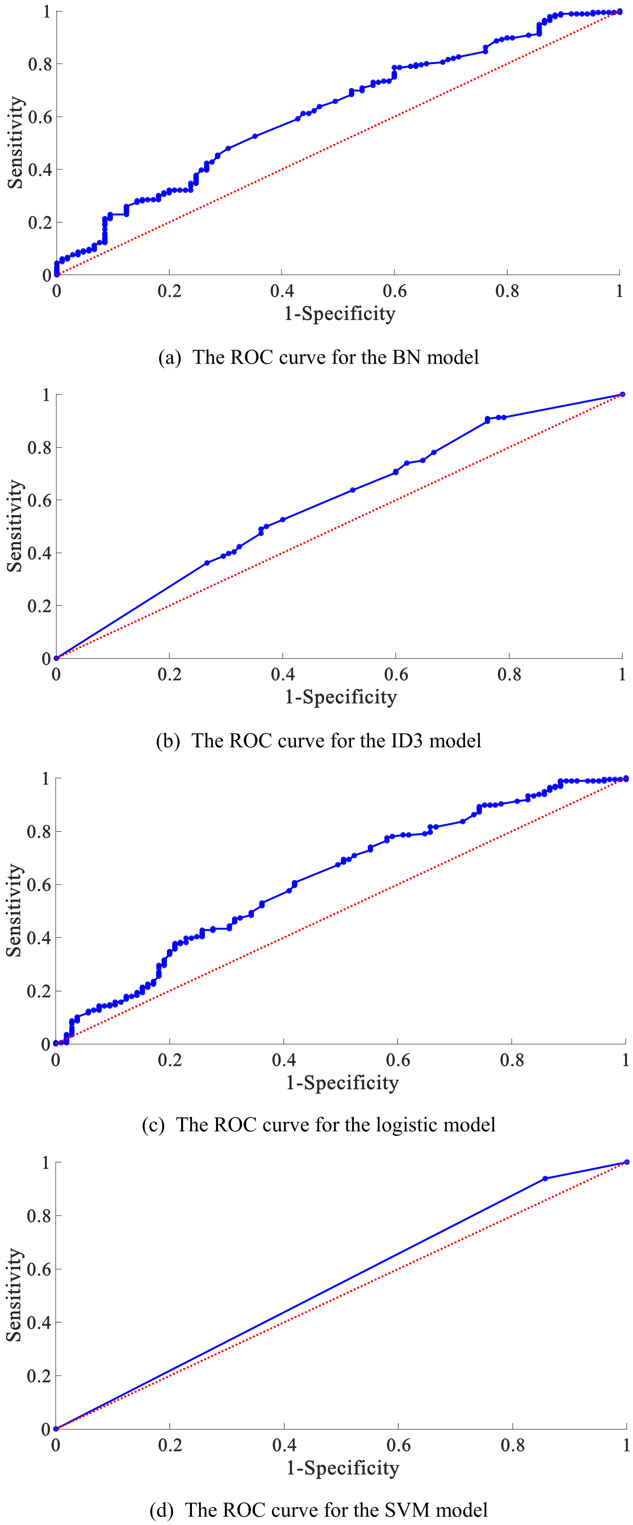
2) DETAILED ACCURACY

Six detailed indicators are listed in Table 2 to evaluate the classification performance of the four algorithms. The results indicated that the four algorithms were able to classify the black spots. Compared with the other three algorithms, The BNs achieved the best precision (0.650) and accuracy (0.668). The highest recall (0.671) and F-measure (0.617) were obtained by the ID3 algorithm. Additionally, the sensitivity of the BNs showed that 66.8% of the cases observed to be black spots were also predicted to be black spots. The highest sensitivity and specificity were for the ID3 algorithm. However, it yielded nine unclassified instances, which led to a lower accuracy compared with that of the BN algorithm.

TABLE 2. Accuracy of the four algorithms.

	TP Rate	FP Rate	Precision	Recall	F-Measure	Accuracy
BNs	0.668	0.580	**0.650**	0.668	0.590	**0.668**
ID3	**0.671**	**0.555**	0.642	**0.671**	**0.617**	0.651
Logistic	0.661	0.579	0.631	0.661	0.589	0.661
SVM	0.661	0.579	0.631	0.661	0.589	0.661

Fig.9 shows an intuitive comparison of the four algorithms. Ideally, the higher the ratio of the TP rate to the FP rate and the values of the other four indicators are (precision, recall, F-measure and ROC area), the better the classification method is. In this study, six evaluation indicators were compared. In terms of the recall, F-measure, and ratio of the



(a) The ROC curve for the BN model

(b) The ROC curve for the ID3 model

(c) The ROC curve for the logistic model

(d) The ROC curve for the SVM model

FIGURE 7. The ROC curves for the four models. (a) The ROC curve for the BN model. (b) The ROC curve for the ID3 model. (c) The ROC curve for the logistic model. (d) The ROC curve for the SVM model.

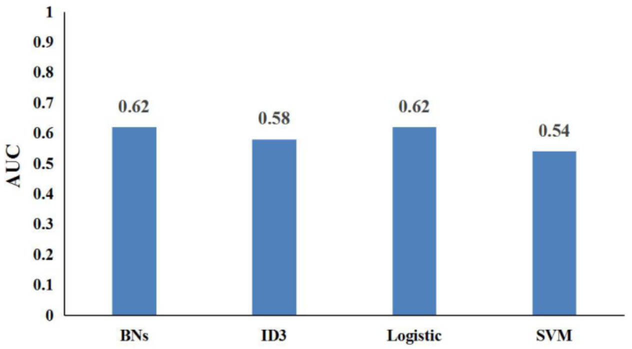


FIGURE 8. AUC for the four models.

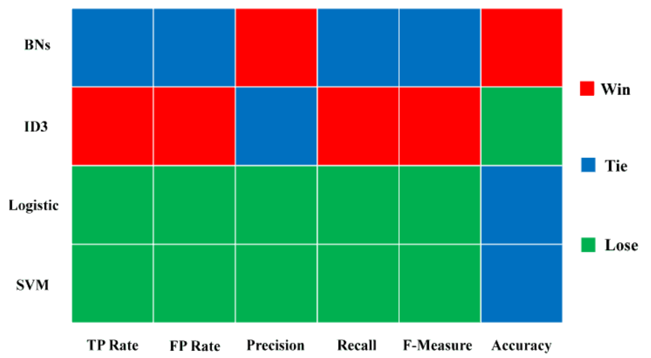


FIGURE 9. Accuracy comparison of the four algorithms.

TABLE 3. Correlation of the variables.

Cor.	1.HOL	2.DAY	3.TIM	4.TRH	5.ACLT	6.ACT	7.WEA	8.RESP	9.BLS
1	1.000	**.093****	.045	.055	.055	**.095****	.001	.060	.034
2		1.000	.006	.043	.011	.022	.008	.011	.005
3			1.000	**.197****	.038	.020	.037	.032	**.104****
4				1.000	.064*	.063*	.051	.066*	.008
5					1.000	**.203****	.017	**.132****	**.266****
6						1.000	**.075***	**.799****	**.143****
7							1.000	.044	.029
8								1.000	**.083****
9									1.000

Cor.	1.HOL	2.DAY	3.TIM	4.TRH
BLS	-0.034	0.005	**-0.104****	-0.008
Cor.	5.ACLT	6.ACT	7.WEA	8.RESP
BLS	**-0.266****	**-0.143****	0.029	**-0.083****

Notes: * Correlation is significant at the 0.05 level; ** Correlation is significant at the 0.01 level.

TP rate to the FP rate, the BN have worse performance than the ID3 algorithm but better performance than the logistic and SVM models. In addition, the BN had the highest precision and accuracy. Although the BN were superior in only two of the six indicators, it had no index of loss. The ID3 algorithm was superior to the other methods in terms of the TP rate, FP rate, recall, and F-measure but was the worst in accuracy. Meanwhile, the logistic and SVM methods both lost in five indexes. Overall, the BN were optimal among the four algorithms and were effective enough to be used to identify the accident black spots.

D. THE RESULT OF THE PEARSON CORRELATION COEFFICIENT TEST

The Pearson correlation coefficient was chosen to test the correlation between the eight impact factors and the black spots using SPSS software. As shown in Table 3, the results show that four factors (accident location type, accident type, time, and responsibility) are significantly correlated with traffic accident black spots ($p<0.05$), where a p-value less than 0.05 indicates the statistical significance of the correlation between the two selected variables. However, HOL, DAY, PEP, and WEA are not significantly correlated with black spots.

In addition, Table 3 also records the calculation results of the correlation coefficient between selected variables and black spots. The absolute value of the correlation coefficient of the accident location type test result is the highest in the dataset, which indicates that the accident location type test result is a critical impact factor that indicates the occurrence of black spots. The variables of 3, 6, and 8 are also associated with accident black spots.

A detailed discussion of the four significant correlation variables is given below.

1) ACCIDENT LOCATION TYPE

As shown in Table 3, the value of the correlation coefficient between ACLT and BLS is -0.266. Compared with the other accident location types (road section, and residential area and parking lot), we can easily find that more black spots appeared at the intersections of roads. This finding is consistent with the result of Al-Ghamdi [19], and Savolainen *et al.* [20], who found that the location has a significant impact on traffic accidents and that more accidents occurred at road intersections than at other locations.

2) ACCIDENT TYPE

ACT was found to be significantly related to BLS. The value of the correlation coefficient between them is -0.141, which indicates that a crash between motor vehicles has the greatest impact on the black spots than other accident types. This is consistent with the findings of Yang *et al.* [21], who found that the number of motor vehicles is an important factor in urban road traffic accidents. It can be concluded that the collisions between motor vehicles needs more attention in future research.

3) TIME

In this study, the time of the accident was divided into day (6:00-18:00) and night (18:00-6:00). The value of the correlation coefficient between TIM and BLS is −0.104. In a previous study, it had already been confirmed that compared with the day, visibility is poorer at night, and the drivers' attention is more easily distracted. These are the main causes of traffic accidents at night [22]. This coincides with the results found in this study.

4) RESPONSIBILITY

The results shown in Table 3 indicate that there is a weak correlation between responsibility and black spots. The value of the correlation coefficient only reaches −0.083. However, it is one of the significant factors and proves that the impact of illegal behaviour on black spots is greater than that of non-illegal behaviour. In Lianfeng Middle Road, the main illegal acts that cause accidents are irregular driving operations, speeding, retrograde and fatigued driving. Hordofa *et al.* [23], Zhang *et al.* [24] highlighted the impact of illegal activities, such as speeding and fatigued driving on traffic accidents.

V. DISCUSSION AND CONCLUSION

This study used the Ningbo traffic accident data to identify traffic accident black spots through the application of the improved K-means clustering and BN algorithms. The BN was compared with the ID3, logistic and SVM algorithms. Several indicators were employed to evaluate the performance of the four models. In addition, a correlation analysis among the variables was carried out based on the experimental data. The results show that 1) the improved K-means algorithm can relatively accurately screen out the traffic accident black spots in the Lianfeng Middle Road area; 2) the four models can effectively identify black spots, and the accuracies are all above 0.6; however, the performance of each model is dissimilar. In contrast, the BN is superior to the other three algorithms; 3) the variables that have the greatest impact on the black spot are accident position type, accident type, time, and responsibility.

This research has some limitations. The improved K-means algorithm is not suitable for cases where the distribution of the accident points has no obvious interval. When the distance between two types of cluster centres is very close, the process of screening out the black spots will be complicated. The variables used in the BN were not comprehensive, and there were no detailed subjective accident factors, such as the personal attributes of drivers and driving speed. Therefore, in future work, it will be necessary to study the K-means and BN algorithms deeply and improve the black spot identification model through more development and experiments to conduct follow-up traffic accident studies.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] *Global Status Report on Road Safety*, WHO, Geneva, Switzerland, 2018.

[2] *Chinese Annual Statistics Report on Road Traffic Accidents*, Ministry of Public Security, Beijing, China, 2017.

[3] X. Shen, X.-C. Guo, and J.-M. Song, "Study on road traffic accident black spot identification method," *J. Highway Transp. Res. Develop.*, vol. 20, no. 4, pp. 95–97, Aug. 2003.

[4] L. B. Meuleners, D. Hendrie, A. H. Lee, and M. Legge, "Effectiveness of the black spot programs in western Australia," *Accident Anal. Prevention*, vol. 40, no. 3, pp. 1211–1216, May 2008.

[5] Y. S. Murat and Z. Cakici, "An integration of different computing approaches in traffic safety analysis," *Transp. Res. Procedia*, vol. 22, pp. 265–274, Jan. 2017.

[6] M. Ghadi and Á. Török, "Comparison different black spot identification methods," *Transp. Res. Procedia*, vol. 27, pp. 1105–1112, Jan. 2017.

[7] A. Gupta, R. Ajaykumar, and P. S. N. Merchant, "Automated lane detection by K-means clustering: A machine learning approach," *Electron. Imag.*, vol. 14, pp. 1–6, Feb. 2016.

[8] Y. Zhong and D. Liu, "The application of K-means clustering algorithm based on Hadoop," in *Proc. ICCCBDA*, Jul. 2016, pp. 88–92.

[9] M. Deublein, M. Schubert, B. T. Adey, J. Köhler, and M. H. Faber, "Prediction of road accidents: A Bayesian hierarchical approach," *Accident Anal. Prevention*, vol. 51, pp. 274–291, Mar. 2013.

[10] R. O. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accident Anal. Prevention*, vol. 88, pp. 37–51, Mar. 2016.

[11] A. C. Mbakwe, A. A. Saka, K. Choi, and Y.-J. Lee, "Alternative method of highway traffic safety analysis for developing countries using Delphi technique and Bayesian network," *Accident Anal. Prevention*, vol. 93, pp. 135–146, Aug. 2016.

[12] G. A. Ali and A. Tayfour, "Characteristics and prediction of traffic accident casualties in sudan using statistical modeling and artificial neural networks," *Int. J. Transp. Sci. Technol.*, vol. 1, no. 4, pp. 305–317, Dec. 2012.

[13] Y. Lu, "The research of Identifying highway accident black spots Based on Empirical Bayes method," M.S. thesis, Dept. Transp. Plan. Mgt., Chang'an Univ., Xi'an, China, 2015.

[14] M. A. Dereli and S. Erdogan, "A new model for determining the traffic accident black spots using GIS-aided spatial statistical methods," *Transp. Res. A, Policy Pract.*, vol. 103, pp. 106–117, Sep. 2017.

[15] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Phys. A, Stat. Mech. Appl.*, vol. 517, pp. 29–35, Mar. 2019.

[16] B. Debrabant, U. Halekoh, W. H. Bonat, D. L. Hansen, J. Hjelmborg, and J. Lauritsen, "Identifying traffic accident black spots with Poisson-Tweedie models," *Accident Anal. Prevention*, vol. 111, pp. 147–154, Feb. 2018.

[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Jan. 1967, pp. 281–297.

[18] J. de Oña, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Anal. Prevention*, vol. 43, no. 1, pp. 402–411, 2011.

[19] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Anal. Prevention*, vol. 34, no. 6, pp. 729–741, 2002.

[20] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident Anal. Prevention*, vol. 43, no. 5, pp. 1666–1676, 2011.

[21] S. W. Yang, W. Wang, and R. Li, "Macro factors of urban road traffic accident," *Modern Transp. Technol.*, vol. 14, no. 3, pp. 82–85, Jun. 2017.

[22] J. J. Rolison, S. Regev, S. Moutari, and A. Feeney, "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records," *Accident Anal. Prevention*, vol. 115, pp. 11–24, Jun. 2018.

[23] G. G. Hordofa, S. Assegid, A. Girma, and T. D. Weldemarium, "Prevalence of fatality and associated factors of road traffic accidents among victims reported to Burayu town police stations, between 2010 and 2015, Ethiopia," *J. Transp. Health*, vol. 10, pp. 186–193, Sep. 2018.

[24] Y. Zhang, T. Liu, Q. Bai, W. Shao, and Q. Wang, "New systems-based method to conduct analysis of road traffic accidents," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 54, pp. 96–109, Apr. 2018.