IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 12, 2020, accepted June 11, 2020, date of publication June 15, 2020, date of current version July 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002535

A Spatial Analysis Methodology Based on Lazy Ensembled Adaptive Associative Classifier and GIS For Examining the Influential Factors on Traffic Fatalities

CHONG ZHAI[1], ZHENG LI[2], FEIFENG JIANG[3], JACK J. MA[2], AND ZHERUI XU[4]

[1]Shenzhen Qianhai Bruco Consulting Company Ltd., Shenzhen 518108, China
[2]Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong
[3]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong
[4]Shenzhen Topband Company Ltd., Shenzhen 518108, China

Corresponding author: Zherui Xu (xuzherui0527@outlook.com)

ABSTRACT Analyzing the influential factors of traffic accidents has been a hot topic in city management. Most existing literature in this domain implemented linear based sensitivity analysis in statistics to study the problems. However, the linear assumption limits their model performance and therefore interferes with the detection of influential factors. Recent studies started to use nonlinear machine learning methods to explore the problem. One of the most popular ways is the association rule analysis. Based on the Support and Confidence value, researchers were able to identify the top influential factors. However, (1) the identification of the thresholds for Support and Confidence has not been well solved in related studies. This study, therefore, proposes Lazy ensembled adaptive Associative Classifier to tackle this problem. Besides, (2) most of the existing literature only analyzed the general relationships between the influential factors and the traffic fatality but did not further investigate their spatial connections. Those studies could not answer specific questions like ''which region should be focused more on alcohol control?'', or ''where requires more attention on motorcycle control?''. This study combines the road-based GIS analysis and the results from association rule analysis to spatially analyze the relationships between the impact factors and the traffic fatalities. Specific suggestions on city management and traffic control were proposed thereafter.

INDEX TERMS Association rule analysis, GIS, machine learning, road-based analysis, traffic accident fatality.

I. INTRODUCTION

Road traffic accidents (RTAs) have become a global public health and development problem, killing nearly 1.3 million people and disabling 20-50 million people annually and costing most countries 3% of their gross domestic product [1]. RTAs have been reported as ''the eighth leading cause of death globally''. However, interventions implemented by countries in past years have proved that most traffic crashes are both predictable and preventable [1]. To support the prediction and prevention of RTAs, scholars and governors should have a proper understanding of the influential factors of traffic accidents. Identifying the factors can help better understand the cause-effect behind, and thus help design relevant interventions. Some studies have been conducted to investigate various kinds of influential factors, such as speed, alcohol, helmets, seat-belts, and road infrastructure [2]–[5]. To better model the relationships and evaluate the factor importance, scholars have proposed different kinds of methods to analyze these impact factors. Commonly used methods can be classified into the following groups.

The first group of methods bases on multivariate regression. For example, Zhong-Xiang *et al.* [6] combined Verhulst and multivariate linear regression models to analyze the fatalities of road traffic accidents in China from 2002 to 2011; Girotto *et al.* [2] investigated the relationship between professional experience and traffic accidents or near-miss accidents among truck drivers using multinomial logistic

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

regression. Anastasopoulos *et al.* [3] used the multivariate Tobit regression model to analyze the highway accident-injury-severity rates. However, the linear assumption behind these methods (e.g., $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$) goes against the non-linearity of the influential factors in real-world problems and affects the model performance [7], [8].

The second group of methods relies more on latent class analysis. For example, Depaire *et al.* [9] applied latent class clustering to identify homogenous traffic accident types. Adanu *et al.* [10] investigated the factors that influence the severity of single-vehicle crashes that happen on weekdays and weekends with the latent class logit model. Latent class analysis can capture unobserved heterogeneity by allowing parameters to differ across observations, but it does not account for the possibility of variation within a class as it assumes homogeneous characteristics of the within-class observations [11], [12].

Some researchers applied non-parameter methods, such as the Bayesian network, to reveal the connection between traffic accidents and their influential factors. For example, de Oña *et al.* [13] used Bayesian Networks as well as the latent class clustering method to study 3229 accidents on rural highways in Granada between 2005 and 2008. Theofilatos [4] deployed Bayesian and finite mixture logit models to investigate the accident likelihood and severity on urban arterials, finding that traffic variations had a significant effect on accident occurrence but mixed effects on accident severity. Elvik *et al.* [14] developed a before-after evaluation of road safety to study the impact of a new motorway in Norway with Empirical Bayes. The problem of the Bayesian network is that it is computationally expensive and not effective on high-dimensional datasets [15].

To avoid the shortcomings of the previous methods, scholars have developed artificial intelligence (AI) related algorithms recently. In data analysis, AI-related methods can be divided into machine learning methods and deep learning methods [16], [17]. Deep learning methods have caught lots of attention these years in traffic analysis due to its superb nonlinear modeling ability [18], [19]. However, its black box nature restricts it from analyzing impact factors [20]. Therefore, the current research directions on this problem shift more towards machine learning (ML) methods. Among all the reported ML methods, one typical example is the association rule analysis. It can not only study the cause-effect between one item factor and the target but also investigate the relationships between multiple item factors and the target. For example, Montella *et al.* [5] applied association rule to reveal the characteristics of powered two-wheeler crashes. Xi *et al.* [21] analyzed the level of influence of causational factors for traffic accidents by association rules. Weng *et al.* [22] investigated the crash casualty patterns of the work zones using association rules. However, the identification of thresholds of association rule analysis remains to be a problem. Previous literature usually relied on the experience of researchers to identify the thresholds [23]–[26], which is not a method that can be generalized. Therefore, an association rule mining-based framework with the ability to identify the thresholds is worthy of attention.

Another limitation of the existing studies on the influential factors is that most of them only investigated the overall weights or relationships between the factors and the traffic fatalities. Few of them went further and analyzed the spatial relationships. For example, an impact factor that has been recognized by many studies is driving with alcohol (or drunk driving), but previous studies failed to answer the question that which place in the city should enhance the alcohol control. This answer to the question is the result of what we defined the spatial analysis of the influential factors in this study.

This study proposes a methodology framework based on association rule analysis and road-based GIS analysis to investigate the influential factors that cause traffic fatalities. The methodology integrates Lazy ensembled adaptive Associative Classifier (LeaCA) to optimize the thresholds and uses geographical information system (GIS) to interpret the cause effects spatially. Traffic accident data of Los Angeles are used to test the framework. By providing evidence-based information, our results can help governments identify the major causes of traffic fatalities in Los Angeles and formulate specific policies and legislation.

The rest of this paper is organized as follows. In Section 2, we describe the methodology framework. In Section 3, a case study in Los Angeles is presented. Discussions of results are given in Section 4. Conclusions and limitations of this paper are provided in Section 5.

II. METHODOLOGY FRAMEWORK

The proposed methodology framework is shown in FIGURE 1. It consists of three parts. The first part is data preprocessing. The second part is the model implementation. Association rule analysis was conducted to investigate the relationships between impact factors and traffic fatalities. The lazy associative classifier was proposed to optimize the threshold for support and confidence in association rule mining. The third part is post engineering, including rule mining and road-based GIS analysis.

A. PREPROCESSING

The first part is data preprocessing. The collected raw data usually has some problems, such as missing values, noisy data, and data imbalance. These problems should be addressed before we use the data in the model. The procedures to tackle these problems are typical in machine learning but may vary a bit from problem to problem. More details will be introduced in the case study.

Besides, the features need to be binarized since the association rule mining can only analyze binary data. In this study, there involved four kinds of features, including binary features, categorical features, numerical features, and string features. Binary features do not need any formatting since it can directly fit association rule analysis. Categorical features are transformed into binary features using the one-hot

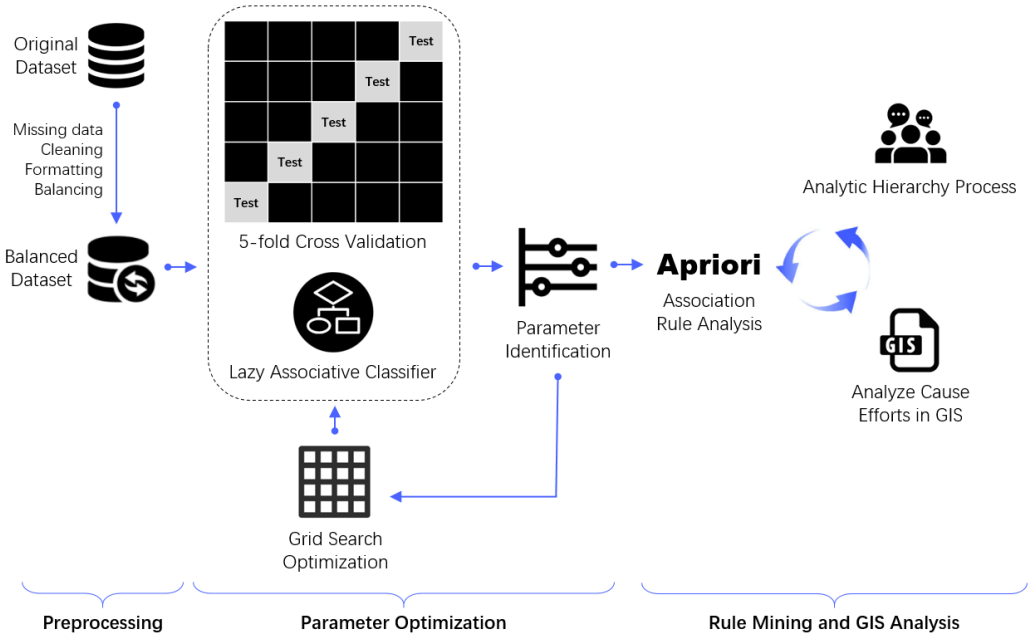

FIGURE 1. Methodology framework.

Accident ID	Victim Role (VR)
001	Driver (D)
002	Passenger (Pa)
003	Pedestrian (Pe)
004	Bicyclist (B)
005	Other (O)

VR-D	VR-Pa	VR-Pe	VR-B	VR-O
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

FIGURE 2. An example of one-hot encoding.

encoding [27]. This technique will generate new binary features to represent each option in the categorical feature. FIGURE 2 presents an example of one-hot encoding.

The formatting of numerical features and string features is a bit complicated. The idea is transforming these features into categorical features first and then modify them into binary features using one-hot encoding. Numerical features may use binning methods to achieve so, while string features are more complicated and the procedures may require much domain knowledge and can vary from problems to problems. More details will be introduced in the case study.

B. MODEL IMPLEMENTATION

1) ASSOCIATION RULES ANALYSIS

Association rules analysis is a rule-based machine learning method for determining the connections between different fields of data. Due to its excellent performance in identifying strong rules in databases, it has been employed in many application areas, such as market basket analysis, web usage mining, and bioinformatics [28]. Association rules mining was firstly introduced by Agrawal *et al.* [29] and can be defined as follows.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of m binary features. Let $D = \{s_1, s_2, \ldots, s_n\}$ be a set of accidents that form the database. Each accident in D has a unique ID and contains a subset of features in I. A *rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of features X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule. In this study, Y has only one indicator of whether the accident is fatal or the victim is dead, while X is the combination of different accident situations and attributes. In this way, the identification of the strong rules in $X \Rightarrow Y$ can help identify the influential factors.

Theoretically, numerous rules can be generated. However, not all rules can provide useful information. Rules which surpass user-specified minimum support and minimum confidence threshold are defined as interesting rules that may reveal valuable knowledge [30]. Support (S) and confidence (C) are two critical criteria in association rules mining. Support determines how often an item appears in the given dataset, and confidence indicates how frequently items in Y appear in transactions that contain X. Their mathematical format can be expressed as Eq. (1) and (2).

$$\text{Support, S}\,(X \to Y) = \frac{\sigma(X \cap Y)}{n} \tag{1}$$

$$\text{Confidence, C}\,(X \to Y) = \frac{\sigma(X \cap Y)}{\sigma(X)} \tag{2}$$

where σ is summation notation.

The identification of the rules and the calculation of its support and confidence can be time-consuming because when the number of features m gets larger, the combination of the features in X can be massive. It is not smart to conduct a brute force procedure to accomplish this. Therefore, scholars proposed the Apriori algorithm to tackle this problem. The algorithm uses a breadth-first search strategy to count the support of feature sets and uses a candidate generation function that exploits the downward closure property of support. The pseudo-code of Apriori is as shown in

ALGORITHM 1. However, in many cases, the efficiency of Apriori is still not satisfactory, especially for long patterns [31]. In this study, since the length of itemsets is less than four in later experiments and the consequent item is fixed as the fatality of the traffic accident, then the rule extraction process is not very complex, and there is no much difference between these choices. Also, the main focus of the methodology is to propose an association rule-based classification model, so we will just go for the most typical and publicly accepted algorithm, Apriori, for rule extraction.

Algorithm 1 The Apriori Algorithm for Generating Candidates of Strong Rules

L_k : frequent feature sets of size k

1. **for** (k =1; $L_k! = \emptyset$; k++)
2. C_{k+1} = candidates generated from L_k;
3. **for** each transaction t in database
4. increment the count of all candidates in C_{k+1} that are contained in t
5. **end**
6. L_{k+1} = candidates in C_{k+1} with min_support
7. **end**
8. **return** $\cup_k L_k$;

It can be inferred from ALGORITHM 1 that the identification of the minimum threshold of support and confidence is one of the prerequisites in association rules analysis. Previous literature usually relied on the experience of the scholars to determine the threshold [30], which, however, is not a method that can be generalized. In this paper, this problem is addressed by integrating a classification algorithm, namely Lazy ensembled adaptive Associative Classifier (LeaAC).

2) LAZY ENSEMBLED ADAPTIVE ASSOCIATIVE CLASSIFIER (LeaAC)

The idea is to build a classification model that has the same target as the association rule mining. For example, this study intends to mine the association rule that will lead to fatal accidents. Then the proposed method will build a classification model based on the boolean features to classify whether an accident is fatal or not. The classification algorithm we developed here is LeaAC, and in this algorithm, Support and Confidence are two parameters. Therefore, the set of parameters that help LeaAC to achieve the highest classification accuracy becomes the optimal value for Support and Confidence. The equation format of this optimization idea is shown in (3), (4), and (5).

$$\{i_1, i_2, \ldots, i_m\} \xrightarrow{rulemining} T \tag{3}$$

$$\{x_1, x_2, \ldots, x_m\} \xrightarrow{classification} T \tag{4}$$

$$\{\mathrm{S}, \mathrm{C}\} = \arg min \sum |T - T'| \tag{5}$$

where T is the target, and is the fatality of the accidents in this study. T' is the predictions of the classification model, x_m are the values of the binary features.

Traditional associative classifier mines all frequent class association rules (CARs) as essential decision-rules [32]. It checks whether each CAR matches the test instance during the testing phase and chooses the first CAR matching the test instance to predict the class. However, it may generate a large number of rules, many of which may be useless, and in some cases, important rules may never be mined [33].

Lazy associative classifier (LAC) overcomes this problem by focusing the rule mining in the given test instance. Instead of creating the classification model during the learning phase using training data, LAC postpones generalization and builds the classification model until a query is given. Although the testing stage can, therefore, be slower, the accuracy can be improved significantly. Also, this study upgrades the labeling process of LCA by introducing adaptive weights for the rules used for classification. The weights are calculated using the information gain in each rule, and the eventual output incorporates the idea from ensemble learning to gather the prediction results of all the rules. We name this algorithm as Lazy Ensembled Adaptive Associative Classifier (LeaAC). The pseudo-code of LeaAC is as shown in ALGORITHM 2.

Algorithm 2 Lazy Ensembled Adaptive Associative Classifier (LeaAC)

D: the set of all n training instances
T: the set of all m test instances
y: the target class (traffic fatality in this study, 1 means fatal, while −1 means non-fatal)

1. **for** each $t_i \epsilon T$ **do**
2. let D_i be the projection of D on features only from t_i
3. let L_i be the set of all rules $\{X \rightarrow y\}$ mined from D_i passing min support and min confidence
4. Calculate the information gain vector G_i of all the rules in L_i
5. Ensemble the results $G_i \cdot \hat{y}_i$ and predict class y_i (positive/negative separation)
6. Insert y_i to Y
7. **Return** Y

C. RULE MINING AND GIS ANALYSIS

After the optimized support and confidence value are obtained, association rules can be extracted using the Apriori algorithm. The rules will then be examined through an analytic hierarchy process (AHP) to determine the real influential factors. AHP is one of the techniques of Multi-Criteria Decision Making (MCDM) to weight and compare a set of elements and then select the best one. Different decision-makers first give out their opinions on the factor weights and factor values, and AHP will integrate their opinions using weighted regression. The top rank factors in the AHP process then become the most appropriate rules. Note that the AHP method relies on the knowledge from domain experts, and their opinions may be subjective to some extent. However,

the problem we target in this study in a complicated real-world city governing problem. The procedure cannot merely be a numerical analysis and avoid opinions from domain experts, and AHP is a scientific tool to collect and integrate the knowledge from experts, while the association rule analysis provides essential preliminary results.

Then, GIS is used to study the spatial relationships between the impact factors and traffic fatality. Traditionally, when plotting the distribution of the factors, scholars may directly use the density plot [34], [35]. However, the density plot of the accidents associated with different factors generally follows the same distribution of the accident density, which makes the density plot less sensitive when studying the spatial relationships between impact factors and traffic fatalities.

In addition, instead of the number of fatal accidents, city managers might be more interested in the fatality rates of accidents. Places with more fatal accidents may simply because they have higher traffic volumes. However, places with higher fatality rates indicate that the place is dangerous, and should be given more attention.

To achieve the spatial analysis of fatality rates, we proposed a road-based analysis in GIS, because traffic accidents are all happened on or near the roads. This study collected all the road data from Los Angeles County GIS Data Portal, and these roads are that used by the US Census Bureau to help locate citizens during its decennial census. The proposed spatial analysis is conducted as follows.

- Map the accident data into the road maps. Since the roads are line features, the accident points cannot directly be joined into the roads. We created polygon buffer zones along both sides of the roads 10 meters, and then map those accidents points into the roads.
- Transfer road features to point features. Although the accidents have been grouped into the roads, the road features are not friendly in visualization. Some roads are long while some are short, so the plotted network can be visually messy and not friendly to analyze. To tackle these, we used points to represent the roads. Each road is transformed into a point, which is located at the central position of the road. We then plot the relevant rates and relationships using those points in GIS.
- Value calculation. After mapping the accident points, we were able to calculate the fatality rates in each road and the relevant accident features, which allows the spatial analysis of these influential factors. More details will be introduced in the case study.

III. AN EXPERIMENTAL CASE IN LOS ANGELES CITY

A. DATA COLLECTION

To validate the proposed methodology, we conducted a case study in Los Angeles city. We choose this city because it is reported to have the highest rate of injury-causing and fatal traffic accidents in the nation [36]. The data was extracted from the open dataset of the Transportation Department of California and the American Highway Control Center (https://dot.ca.gov/). This study focuses on the fatality of the traffic accidents, while the fatality level of an accident is decided by whether there is any victim been killed. Therefore, this study uses the fatality of the victim as the research target. This target can provide more insights from the perspective of the victim to understand fatal accidents. Alongside the ten-year accidents (2003-2012) from the raw data, we obtained 526,123 victims and the information of the related accidents. Based on the fatality of the victims, we obtained 43,668 positive cases (fatal), and 482,455 negative cases (non-fatal).

TABLE 1 presents the detailed features of the dataset. 73 features are divided into 5 groups, representing features about the collisions (28 features), features regarding the victims (8 features), features of the parties that involved in the accidents (11 features), features concerning the time and location of the accidents (19 features) as well as the features related to the environment (7 features). The second column shows the feature abbreviation and description. The third column shows the data type, and the fourth column presents more details of the features. If the feature is categorical, then the number of categories is shown. If the feature is numeric, then the standard deviation is provided [37].

B. DATA PREPROCESSING

1) DATA FORMATTING

The raw data cannot be directly inputted into the methodology framework due to some flaws. Several preprocessing procedures need to be conducted. The first is data formatting. There are 18 numerical features in the dataset, and these features cannot be directly used in association rules analysis. They need to be converted to binary categorical data. This study implemented an equal bin method. This method first ranks the numerical value from the smallest to the largest and then divides the cases into k different groups with the same frequency. The samples in each group share the same categorical value. FIGURE 3 presents an example of formatting a numerical feature into a categorical feature. By using this method, this study transformed the 18 numerical features into categorical features, and k is set as 5 ($k = 5$ provides the highest accuracy in later experiments).

Note that the only string feature in this study refers to the name of the roads, which is useless in this study, so it was excluded from the experiment. After obtaining the categorical

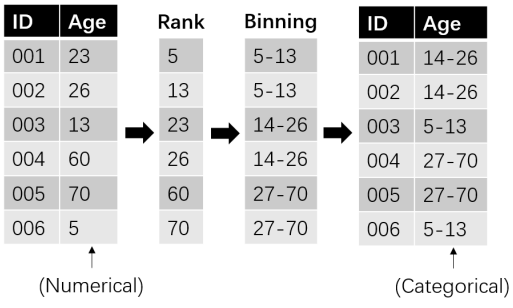

FIGURE 3. An example of formatting a numerical feature into a categorical feature using the equal bin method (k = 3).

TABLE 1. Data description.

	Item abbreviation and Description	Data Type[3]	Detail
Collision related features	**CHPTYPE**：CHP[1] Beat Type	C	12
	VIOLCODE：PCF[2] Violation Code	C	7
	VIOLCAT： PCF Violation Category	C	9
	CRASHTYP：Type of Collision	C	25
	INVOLVE：Motor Vehicle Involved With	C	10
	PED：Pedestrian Action	C	7
	CHPFAULT：CHP Vehicle Type is at fault	C	99
	SHIFT：CHP effect of new 12-hour shifts	C	4
	VIOLSUB：PCF Violation Subsection	C	5
	BEATTYPE：Beat Type	C	9
	PARTIES：Party Count	N	2.28±1.03
	PEDCOL：whether involved a pedestrian	B	/
	BICCOL：whether involved a bicycle	B	/
	MCCOL：whether involved a motorcycle	B	/
	TRUCKCOL：whether involved a big truck	B	/
	ETOH：whether involved drinking party	B	/
	STFAULT：indicates who is at fault	C	15
	HITRUN：Hit and Run	C	3
	PCF：Primary Collision Factor	C	6
	RIGHTWAY：Control Device	C	5
	NOTPRIV：whether on private property	B	/
	DIRECT：Direction of the offset distance	C	4
	INTERSECT_：whether at an intersection	B	/
	KILLED：counts of victims with 1-degree injury	N	0.02±0.17
	INJURED: counts of victims with 2,3,4 of injury	N	1.89±1.55
	CRASHSEV: the severity level of the collision	N	3.56± 0.64
	BEATNUMB: Beat Number	C	999
	VIOL: PCF Violation	C	999
Victim related features	**VTYPE**：victim role	C	6
	VSEX：victim sex	C	2
	VAge：victim age	N	0-100
	Vseat：victim seating position	C	9
	Vsafety1：victim safety equipment	C	26
	Vsafety2：victim safety equement2	C	26
	Vejected：victim Ejected	C	3
Parties involved	**Ptype**：party type	C	5

TABLE 1. *(continued)* **Data description.**

	Atfault: if party was at fault	B	/
	Psex: party sex	C	2
	Page: party age	N	0-99
	Psober: indicates drink influence	C	6
	Pdruy: indicates party drug influence	C	4
	Psafety1: party safety equipment	C	26
	Psafety2: party safety equipment	C	26
	Insured: Financial responsibility	C	4
	Cell: includes party cell if in use	C	5
	Vehyear: the model year of the party's vehicle	N	2000-2012
Temporal and spatial features	**YEAR_:** Collision Year:	C	10
	MONTH_: The month of the year	C	12
	DAYWEEK: The Day of Week	C	7
	TIMECAT: 3-hour categories time	C	9
	DATE_: the date when the collision occurred	N	/
	PROCDATE: Date the record was processed	N	/
	TIME_: the time (24-hour time)	N	/
	LAPDDIV: City Division LAPD	N	/
	STATEHW: whether on a state highway	B	/
	POINT_X: The longitude of the geocoded location	N	/
	POINT_Y: The latitude of the geocoded location	N	/
	JURDIST: Reporting District	N	/
	DISTANCE: Offset distance from a secondary road	N	277.26± 3040.27
	LOCATION: the location code PRIMARYRD	N	/
	JURIS: Jurisdiction	N	4170.20± 3469.51
	POSTMILE: markers indicate the distance a route travels through individual counties	N	5.34±11.91
	SECONDRD: A secondary reference road	S	/
	SPECIAL: Special Condition	C	7
	RAMP: Ramp Intersection	C	8
Environmental features	**WEATHER:** the weather condition	C	8
	WEATHER2: the additional weather condition	C	8
	LIGHTING: lighting condition	C	6
	POP: Population level	C	10
	ROADSURF: Road Surface	C	5
	CHPRDTYP: CHP Road Type	C	9
	RDCOND1: Road Condition 1	C	9

1: CHP: California Highway Patrol
2: PCF: Primary Collision Factor
3: C: Categorical B: Binary N: Numeric S: String

features, we transformed them into binary features using the one-hot encoding methods introduced in the methodology section. After these steps, the data dimension of this experiment is expanded to 682.

2) DATA CLEANING

Besides formatting the data into a model friendly manner, the noisy data need to be excluded. Data cleaning can help reduce the calculation complexity and better interpret the relationships [38], [39]. In this study, we removed two kinds of noisy data, including redundant features and high correlational features. Redundant features describe useless information for mining the influential features on the victim fatalities, so they are excluded from the experiment [40]. For example, features such as ''POING_X'', ''POINT_Y'', and ''LAPDDIV'' describe the spatial coordinates and the jurisdictional information, which are not the causes of traffic fatalities, and therefore, they are deleted.

High correlational features mean some features are too similar to each other, and the existence of these features provides limited additional information for data mining but increases the complexity, and therefore, they should be excluded as well. Since the features in this study have already been transformed into binary features, the Pearson correlation is not available for measuring the correlation. Therefore, Spearman correlation is used in this experiment. The difference between these two measures is that Pearson uses numerical values, while Spearman uses rank values. For a binary feature, positive values rank the first, while negative values rank the second. This study excluded one feature in each pair that has an absolute correlation higher than 0.9. The remained one has a higher correlation with the target (victim fatality), while the deleted one has a lower value. After these two steps of data cleaning, the data dimension drops to 399.

3) NEGATIVE ASSOCIATION RULES

Traditional association rule analysis can only discover positive rules because when calculating support and confidence, it will neglect the negative class. However, the negative classes can sometimes provide valuable insights [41], [42]. For example, according to the results in this study, whether the victim has insurance can influence the fatality rate a lot. However, it is not the rule ''the victim has insurance lead to a fatal accident'' is a strong rule, but the reverse, ''the victim has no insurance lead to a fatal accident'' a strong one.

To identify these negative but strong rules, we generated a set of negative features by reversing the positive and negative classes in each feature. The newly created features have a -1 correlation with the original features. After this step, the feature dimension increases to 798.

4) DATA BALANCING AND CROSS-VALIDATION

Another problem that exists in this study is the imbalance. As introduced in data collection, this experiment has 43,668 positive cases and 482,455 negative cases. The dataset is very imbalanced. The rate between positive cases and negative cases is around 1:11.

Traditionally, scholars would use either under-sampling or oversampling to address this problem. However, with such a large imbalance rate, oversampling can easily cause overfitting [43], while under-sampling can miss a large proportion of the data. Therefore, this study proposes a combined strategy to address this issue. This strategy will divide the negative cases into 11 segments without replacement, and then conduct the modeling procedure 11 times. Each time the positive cases will be combined into one segment of the negative cases to form a dataset for modeling and calculation. The averaged results of these 11 runs give the eventual results.

Also note that thanks to the 11-run strategy, there is no need for cross-validation in this study. The averaged performance of the traditional 3/7 testing/training partition of these 11 runs can already provide stable and reliable results for both classifications using LeaCA and association rule mining. Note that the random seed in each run is different, so the positive cases in these 11 runs are also different.

IV. RESULTS AND DISCUSSION

A. IDENTIFICATION OF THE OPTIMAL THRESHOLDS FOR SUPPORT AND CONFIDENCE

This experiment targets at studying the influential factors on the victim fatalities using association rule analysis. Support and Confidence are two criteria to filter out numerically strong rules. One problem in the existing literature is that they cannot identify a set of proper thresholds for these two criteria. This study proposes the implementation of LeaCA models to address this gap. The idea is using the influential features as the variables and the victim fatalities as the target to build binary classification models with the data balanced. Support and Confidence are two critical parameters in this model, so the model that provides the best classification performance defines the optimal thresholds for Support and Confidence.

After preprocessing, the dataset can be fed into the classification model built by the LeaCA algorithm. Besides Support and Confidence, there is another parameter that affects the performance of LeaCA a lot. That is the number of maximum items in a rule and is marked as I_{max}. Therefore, in order to identify the best set of Support and Confidence, this study optimizes these three parameters together.

This experiment explored the model performance when $I_{max} = \{2, 3, 4\}$. Note that when $I_{max} = 2$, the rules generated by the Apriori algorithm only has two items, which consist of one antecedent and one consequent (such as A $\Rightarrow$ B), while when $I_{max} = 3$, three-item rules such as A, C$\}$ $\Rightarrow$ B, can be generated.

After some tests, we found when $I_{max} = 4$, the training time of the model is too long to be acceptable, and the accuracy drops significantly, so we explored when $I_{max} = \{2, 3\}$. FIGURE 4 and FIGURE 5 present the optimization procedures of Support and Confidence with different I_{max}.

TABLE 2. Classification accuracy of the four algorithms. The results are presented using the mean ± standard deviation format of the 11 runs.

Algorithm	Accuracy
LeaAC	82.31%±0.67%
MLR	81.60%±0.68%
LR	78.83%±0.72%
NB	81.92%±0.53%

It is discovered that when $I_{max} = 2$, the highest modeling accuracy is 82.31%, and when $I_{max} = 3$, the highest modeling accuracy becomes 77.78%. Therefore, I_{max} is set as 2, and the best performance model is given by the parameter set with Support=0.04 and Confidence =0.74. As a result, this set of parameters is set as the threshold for association rule mining in this study. We think the reason why $I_{max} = 2$ outperforms $I_{max} = \{3, 4\}$ is because longer rules contained more constrains and has higher risks of overfitting.

Besides, to further verify the effectiveness of the proposed methodology, we added a comparison with three other commonly used methods on modeling the feature weights. They are multiple linear regression (MLR), logistic regression (LR), Naive Bayes. MLR and Naïve Bayes are the most typical algorithms mentioned and the first and third group of methods in the introduction, while logistic regression is the most commonly used nonlinear regression methods in the industry. For the latent class analyses mentioned in the introduction, since they are unsupervised learning methods and do not support regression, we did not pick them for comparison here. TABLE 2 presents the results of the comparison. The proposed LeaAC method has the highest modeling accuracy. This performance, on another angle, supported the priority of the proposed method.

B. STRONG RULES ON VICTIM FATALITY

After obtaining the optimal Support and Confidence, we applied the thresholds on the 11 datasets generated in the balancing step. To reduce the impact of data variance, the rules that survive in all these 11 runs are extracted as the strong rules. This resulted in 69 strong rules in this study.

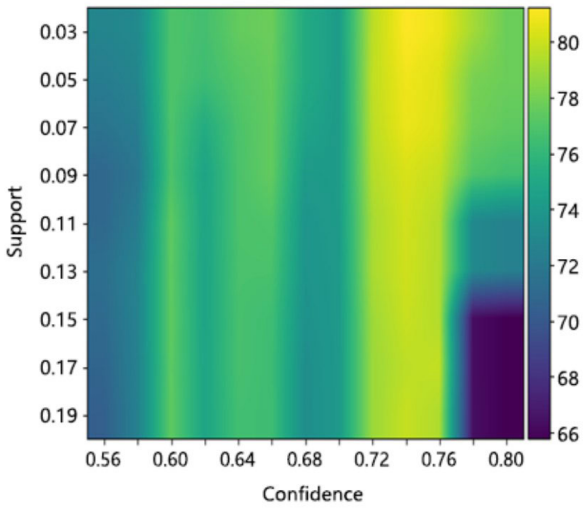

FIGURE 4. Parameter optimization when $I_{max} = 2$.

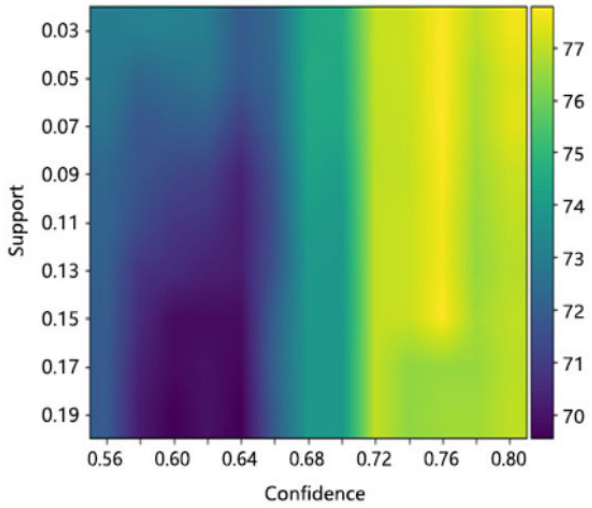

FIGURE 5. Parameter optimization when $I_{max} = 3$.

However, not all the 69 rules are of practical use, and some of them become "strong" may because of data variance. Therefore, this study conducted the Analytic Hierarchy Process (AHP) method to identify practical strong rules further. AHP is a decision-making process that will collect opinions from domain experts first to decide the weights of the decision-making factors and then decided the score of a candidate at each factor [44], [45]. For example, we defined three criteria to identify strong rules. Besides Support and Confidence, we added "practicality" that also ranges from [0,1] to measure the practicality of a rule. This "practicality" is to collected the domain experts' opinions on the practicality of a rule in the questionnaire. An expert should first mark the practicality score of a rule (P in Equation IV-C) and then provide his opinion on the decision-making weights of these three criteria (W_S, W_C, W_P, in Equation IV-C). Support and Confidence already have calculated values, so the expert does not need to score for them. The following equation then gives the final score of a rule.

$$\text{Score} = \text{S} \cdot W_S + C \cdot W_C + P \cdot W_P \tag{6}$$

where S and C are Support and Confidence value, P is the practicality score, and W is the weights provided by the experts.

The questionnaire is sent out to fifty scholars that have related publications on machine learning or statistics in accident research. Sixteen of them replied, and eleven of them completed the questionnaire. We gathered their opinions, averaged their weights in Equation IV-C, and calculated the final score of all the 69 rules. TABLE 3 lists the top 10 rules with the highest score. These are recognized as the strong practical rules in this study.

TABLE 3. Top 10 features.

No.	antecedents	support	confidence	description
1	(VSAFETY2_W)	0.0785	0.8800	Victims are motorcycle drivers with helmets
2	(VEJECTED_1)	0.1121	0.8722	Victims are fully ejected from their seats
3	(VIOLCAT_11)	0.0678	0.8636	Pedestrian violation
4	(VTYPE_3)	0.1492	0.8461	Victims are pedestrians
5	(PSOBER_B)	0.06035	0.8366	Parties involved are under alcohol influence
6	(notINSURED_Y)	0.3321	0.7868	Proof of insurance is not obtained or insurance is not applicable
7	(notPSAFETY2_G)	0.3453	0.7829	Parties involved don't use lap/shoulder belt
8	(INSURED_N)	0.0835	0.7826	Proof of insurance is not obtained
9	(PSAFETY1_P)	0.0860	0.7749	Party safety equipment is not required
10	(notVSAFETY2_G)	0.3560	0.7530	Victims don't use lap/shoulder belt

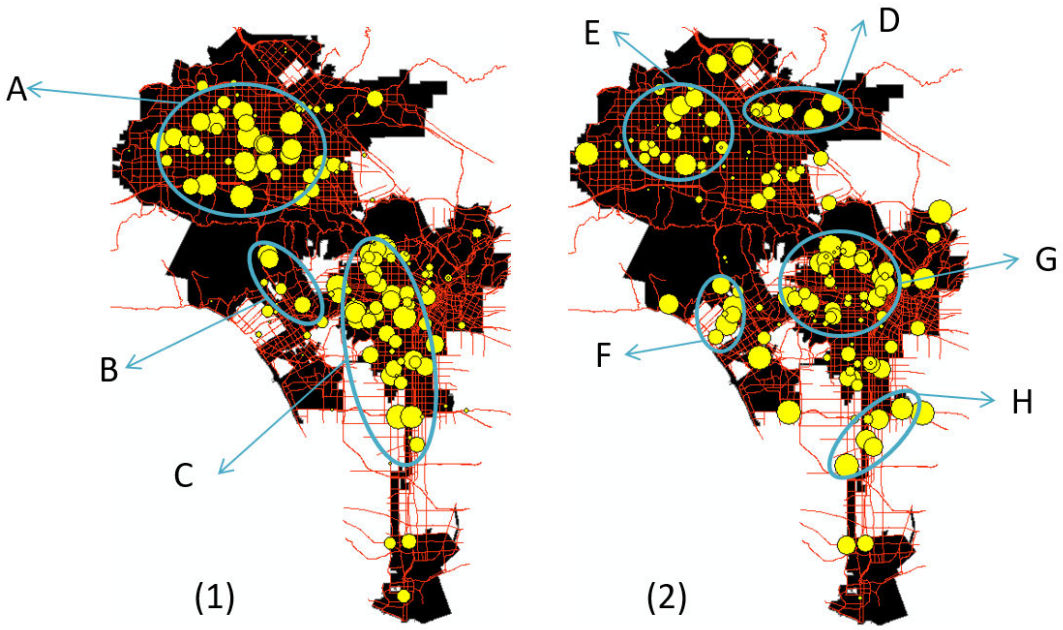

FIGURE 6. (1) The density distribution of the traffic accidents; (2) the distribution of the fatality rates in LA.

C. DISCUSSIONS AND SPATIAL RELATIONSHIPS

It can be seen from TABLE 3 that some of the rules have similar meanings. For example, VIOLCAT_11 and VTYPE_3 are both talking about pedestrians. notINSURED_Y and INSURED_N both mean in the accidents, proof of insurance is not obtained or the insurance is not applicable. So, in this discussion section, we combined the discussion of the rules with similar meanings and got six major influential factors, including VSAFETY2_W, VEJECTED_1, VTYPE_3, PSOBER_B, notINSURED_Y, and notVSAFETY2_G.

These features have been partly discussed by previous literature [4], [10], [22], [46]. All of them exhibit explicit threats that lead to fatal accidents. For example, VSAFETY2_W and VTYPE_3 are talking about two typical Vulnerable Road Users (VRUs), which are motorcycle drivers and pedestrians.

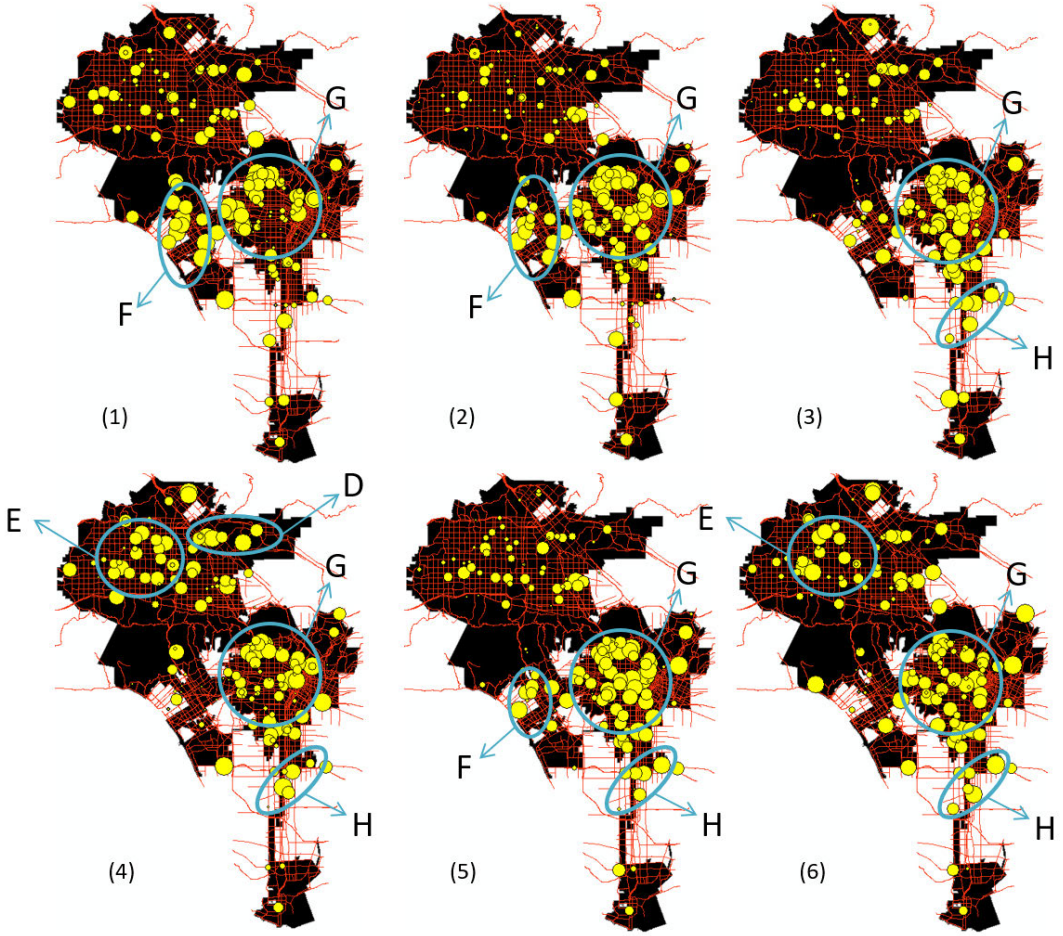

FIGURE 7. Percentage distribution of (1) motorcycle accidents; (2) victims ejected from vehicles; (3) pedestrians involved accidents; (4) alcohol-impaired accidents; (5) victims or parties that do not have insurance; (6) victims that do not use seatbelts.

Studies have shown that these VRUs have five times higher fatality rates than typical car-car accidents [46] because they have no protection equipment such as seat belts or airbags.

VEJECTED_1 refers to the victims that are fully ejected from their seats. This relationship may have two situations. First is that the victim may not have fastened the seatbelt during a severe accident, which is also part of the situation described by the feature notVSAFETY2_G. This apparently can lead to higher fatality rates. The second situation may go back to motorcycle accidents. The drivers or the victims on a motorcycle does not have seatbelts and can be easily ejected from their seats.

PSOBER_B refers to the accidents involved with alcohol, which has been a well-known killer in traffic accidents. Although governments have tried different policies and strategies, drunk driving is still causing many traffic fatalities. This experiment will later point out where should the government focuses more when controlling drunk driving.

notINSURED_Y is describing the group of victims and parties that do not have insurance. The cause-effect behind this feature may result from two aspects. The first may refer to those from low-income families or under the poverty level. They are not willing to buy insurance due to economic issues, and therefore cannot receive proper treatments after a car accident. The fatality rate is then increased. The second aspect may result from those who are not fully aware of the importance of insurance and do not want to waste money on that. This group of people may have limited education, and they may also have a weak sense of traffic rules or proper driving behaviors. These factors potentially lead to higher fatality rates.

To spatially analyze the relationship between the six major influential factors and the fatality rate, the road-based GIS analysis introduced in the methodology section is utilized. FIGURE 6 presents the distribution of traffic accidents and the victim fatality rate. The red line represents the road network, and the yellow cycle represents the density of traffic accidents or the victim fatality rate. Larger cycle means denser accidents or higher rates. It can be observed from the accident distributions that traffic accidents mainly scatter in areas A, B, and C. This might be caused by the dense population and the large traffic volume there. The distribution

of the victim fatality rate shows that areas D to H are more dangerous because they have higher fatality rates than other areas. As a result, instead of discussing the phenomenon behind the high density in areas A to C, this study is more interested in analyzing the influential factor behind the area with high fatality rates (D to H). To achieve this, we analyzed and plotted the percentage distributions of the accidents related to the six features in FIGURE 7 [47].

FIGURE 7 (1) indicates that the percentage of motorcycle accidents are quite high in area F and G. Therefore, to better control and reduce the fatality rate in these two areas, the government is suggested to put more constraints on the motorcycle driving there, such as speed control, forbidding motorcycle in bad weather. FIGURE 7 (2) shows that the percentage of ejected-from-seat accidents are higher in area F and G. The distributions in FIGURE 7 (2) are quite close to FIGURE 7 (1). We might guess that most of the ejected-from-seat accidents relate to motorcycle accidents.

FIGURE 7 (3) reflects that the high fatality rates in areas G and H may be caused by the high rates of pedestrian accidents there. Therefore, the government should consider enhancing pedestrian safety in these areas. For example, design more pedestrian overpasses and underpasses, build more pedestrian guardrails.

FIGURE 7 (4) is the distribution of the percentage of accidents involved with alcohol. It seems that most of the dangerous places involve a high percentage of alcohol-impaired driving, such as area D, E, G, and H. Although it has been a tough task to control drunk driving all over the country for many years, the LA government should know that D, E, G, and H, these four areas should be their focuses.

The percentage distribution of the victims that do not have insurances is shown in FIGURE 7 (5). According to the analysis in previous contents, we suggest the government may enhance the management of compulsory insurance in areas F, G, and H. Also, proper financial support on the insurance in those areas can be considered.

The last figure in FIGURE 7 refers to the percentage of victims not using seatbelts. Therefore, the government may consider increasing the penalty for not wearing seatbelts in area E, G, and H, or investing in AI-empowered video surveillance on seatbelts to strengthen the management.

To sum up, this section discussed the identified influential factors for traffic fatality in LA. Through the road-based spatial analysis in GIS, we provided several suggestions to the government on improving traffic safety. Note that these suggestions are only the results from numerical studies. The real cause effects of the relationships and the effectiveness of these suggestions require further research to verify.

V. CONCLUSION

This paper studied the relationships between fatal traffic accidents and their influential factors in Los Angeles during ten years, using association rule analysis and Geographical Information System (GIS). The problem of determining the minimum thresholds of support and confidence in association rules mining is addressed by applying Lazy Ensembled Adaptive Associative Classifier (LeaAC). Spatial analysis of the relationship between the influential factors and the locations is conducted with the help of GIS. The contributions of this study are as follows:

- The proposed methodology can not only numerically identify the most critical rules on traffic fatality, but also spatially analyze the relationships between the features and fatality rates. This method is expected to be applicable in other cities or regions, as well.
- The LeaAC model addressed the threshold problem in association rule mining, which is viewed as an advanced machine learning method for analyzing influential factors.
- The case study in LA uncovered six important influential factors on traffic fatality. The road-based analysis in GIS provided several actionable suggestions to the government.

On the other hand, this study has limitations. Due to data availability, we only tested the traffic accidents in Los Angeles and did not examine the method performance in other cities and countries. Also, the data used in this study is from 2003 to 2012, which did not reveal the situations in recent years. Future studies can be extended to address these gaps and validate the proposed method in other accident datasets.

REFERENCES

[1] WHO. *Who | Global Status Report on Road Safety 2015*. Accessed: Aug. 10, 2018. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

[2] E. Girotto, S. M. de Andrade, A. D. González, and A. E. Mesas, "Professional experience and traffic accidents/near-miss accidents among truck drivers," *Accident Anal. Prevention*, vol. 95, pp. 299–304, Oct. 2016, doi: 10.1016/j.aap.2016.07.004.

[3] P. C. Anastasopoulos, V. N. Shankar, J. E. Haddock, and F. L. Mannering, "A multivariate tobit analysis of highway accident-injury-severity rates," *Accident Anal. Prevention*, vol. 45, pp. 110–119, Mar. 2012, doi: 10.1016/j.aap.2011.11.006.

[4] A. Theofilatos, "Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials," *J. Saf. Res.*, vol. 61, pp. 9–21, Jun. 2017, doi: 10.1016/j.jsr.2017.02.003.

[5] A. Montella, M. Aria, A. D'Ambrosio, and F. Mauriello, "Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery," *Accident Anal. Prevention*, vol. 49, pp. 58–72, Nov. 2012, doi: 10.1016/j.aap.2011.04.025.

[6] F. Zhong-Xiang, L. Shi-Sheng, Z. Wei-Hua, and Z. Nan-Nan, "Combined prediction model of death toll for road traffic accidents based on independent and dependent variables," *Comput. Intell. Neurosci.*, vol. 2014, Dec. 2014, Art. no. 103196. Accessed: Jul. 7, 2018. [Online]. Available: https://www.hindawi.com/journals/cin/2014/103196/

[7] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, and Z. Xu, "Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques," *Water Res.*, vol. 170, Mar. 2020, Art. no. 115350, doi: 10.1016/j.watres.2019.115350.

[8] J. Ma, Z. Li, J. C. P. Cheng, Y. Ding, C. Lin, and Z. Xu, "Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network," *Sci. Total Environ.*, vol. 705, Feb. 2020, Art. no. 135771, doi: 10.1016/j.scitotenv.2019.135771.

[9] B. Depaire, G. Wets, and K. Vanhoof, "Traffic accident segmentation by means of latent class clustering," *Accident Anal. Prevention*, vol. 40, no. 4, pp. 1257–1266, Jul. 2008, doi: 10.1016/j.aap.2008.01.007.

[10] E. K. Adanu, A. Hainen, and S. Jones, "Latent class analysis of factors that influence weekday and weekend single-vehicle crash severities," *Accident Anal. Prevention*, vol. 113, pp. 187–192, Apr. 2018, doi: 10.1016/j.aap.2018.01.035.

[11] S. T. Lanza and B. L. Rhoades, "Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment," *Prevention Sci.*, vol. 14, no. 2, pp. 157–168, Apr. 2013, doi: 10.1007/s11121-011-0201-1.
[12] J. Ma, J. C. P. Cheng, F. Jiang, W. Chen, and J. Zhang, "Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques," *Land Use Policy*, vol. 94, May 2020, Art. no. 104537, doi: 10.1016/j.landusepol.2020.104537.
[13] J. de Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks," *Accident Anal. Prevention*, vol. 51, pp. 1–10, Mar. 2013, doi: 10.1016/j.aap.2012.10.016.
[14] R. Elvik, H. Ulstein, K. Wifstad, R. S. Syrstad, A. R. Seeberg, M. U. Gulbrandsen, and M. Welde, "An empirical Bayes before-after evaluation of road safety effects of a new motorway in Norway," *Accident Anal. Prevention*, vol. 108, pp. 285–296, Nov. 2017, doi: 10.1016/j.aap.2017.09.014.
[15] *Advantages and Disadvantages of Bayesian Learning Machine Learning (Theory)*. Accessed: Jul. 7, 2018. [Online]. Available: http://hunch.net/?p=65
[16] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, Y. Tan, V. J. L. Gan, and Z. Wan, "Identification of high impact factors of air quality on a national scale using big data and machine learning techniques," *J. Cleaner Prod.*, vol. 244, Jan. 2020, Art. no. 118955, doi: 10.1016/j.jclepro.2019.118955.
[17] J. Ma, J. C. P. Cheng, F. Jiang, V. J. L. Gan, M. Wang, and C. Zhai, "Real-time detection of wildfire risk caused by powerline vegetation faults using advanced machine learning techniques," *Adv. Eng. Informat.*, vol. 44, Apr. 2020, Art. no. 101070, doi: 10.1016/j.aei.2020.101070.
[18] Y. Ding, Z. Li, C. Zhang, and J. Ma, "Prediction of ambient PM2.5 concentrations using a correlation filtered spatial-temporal long short-term memory model," *Appl. Sci.*, vol. 10, no. 1, p. 14, Dec. 2019, doi: 10.3390/app10010014.
[19] F. Jiang, K. K. R. Yuen, and E. W. M. Lee, "A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions," *Accident Anal. Prevention*, vol. 141, Jun. 2020, Art. no. 105520, doi: 10.1016/j.aap.2020.105520.
[20] J. Ma and J. C. P. Cheng, "Identifying the influential features on the regional energy use intensity of residential buildings based on random forests," *Appl. Energy*, vol. 183, pp. 193–201, Dec. 2016, doi: 10.1016/j.apenergy.2016.08.096.
[21] J. Xi, Z. Zhao, W. Li, and Q. Wang, "A traffic accident causation analysis method based on AHP-*apriori*," *Procedia Eng.*, vol. 137, pp. 680–687, 2016, doi: 10.1016/j.proeng.2016.01.305.
[22] J. Weng, J.-Z. Zhu, X. Yan, and Z. Liu, "Investigation of work zone crash casualty patterns using association rules," *Accident Anal. Prevention*, vol. 92, pp. 43–52, Jul. 2016, doi: 10.1016/j.aap.2016.03.017.
[23] A. Pande and M. Abdel-Aty, "Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool," *Saf. Sci.*, vol. 47, no. 1, pp. 145–154, Jan. 2009, doi: 10.1016/j.ssci.2007.12.001.
[24] P. Fournier-Viger. (May 11, 2013). How to Auto-Adjust the Minimum Support Threshold According to the Data Size. The Data Mining Blog. Accessed: Aug. 10, 2018. [Online]. Available: http://data-mining.philippe-fournier-viger.com/how-to-auto-adjust-the-minimum-support-threshold-according-to-the-data-size/
[25] J. Ma and J. C. P. Cheng, "Identification of the numerical patterns behind the leading counties in the U.S. local green building markets using data mining," *J. Cleaner Prod.*, vol. 151, pp. 406–418, May 2017, doi: 10.1016/j.jclepro.2017.03.083.
[26] M. A. Jun and J. C. P. Cheng, "Selection of target LEED credits based on project information and climatic factors using data mining techniques," *Adv. Eng. Informat.*, vol. 32, pp. 224–236, Apr. 2017, doi: 10.1016/j.aei.2017.03.004.
[27] V. J. L. Gan, I. M. C. Lo, J. Ma, K. T. Tse, J. C. P. Cheng, and C. M. Chan, "Simulation optimisation towards energy efficient green buildings: Current status and future trends," *J. Cleaner Prod.*, vol. 254, May 2020, Art. no. 120012, doi: 10.1016/j.jclepro.2020.120012.
[28] (Aug. 11, 2018). *Association Rule Learning*. Accessed: Aug. 17, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Association_rule_learning&oldid=854462239
[29] R. Agrawal, T. Imielinski, A. Swami, H. Road, and S. Jose, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1993, p. 10.
[30] M. Doostan and B. H. Chowdhury, "Power distribution system fault cause analysis by using association rule mining," *Electr. Power Syst. Res.*, vol. 152, pp. 140–147, Nov. 2017, doi: 10.1016/j.epsr.2017.07.005.
[31] J. Ma and J. C. P. Cheng, "Data-driven study on the achievement of LEED credits using percentage of average score and association rule analysis," *Building Environ.*, vol. 98, pp. 121–132, Mar. 2016, doi: 10.1016/j.buildenv.2016.01.005.
[32] J. C. P. Cheng and L. J. Ma, "A data-driven study of important climate factors on the achievement of LEED-EB credits," *Building Environ.*, vol. 90, pp. 232–244, Aug. 2015, doi: 10.1016/j.buildenv.2014.11.029.
[33] A. Veloso, W. Meira, Jr., and M. J. Zaki, "Lazy associative classification," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 645–654, doi: 10.1109/ICDM.2006.96.
[34] J. Ma, J. C. P. Cheng, Y. Ding, C. Lin, F. Jiang, M. Wang, and C. Zhai, "Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series," *Adv. Eng. Informat.*, vol. 44, Apr. 2020, Art. no. 101092, doi: 10.1016/j.aei.2020.101092.
[35] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, and Z. Wan, "A temporal-spatial interpolation and extrapolation method based on geographic long short-term memory neural network for PM2.5," *J. Cleaner Prod.*, vol. 237, Nov. 2019, Art. no. 117729, doi: 10.1016/j.jclepro.2019.117729.
[36] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep learning: A generic approach for extreme condition traffic forecasting," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 777–785.
[37] J. Ma and J. C. P. Cheng, "Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology," *Appl. Energy*, vol. 183, pp. 182–192, Dec. 2016, doi: 10.1016/j.apenergy.2016.08.079.
[38] J. Ma, J. C. P. Cheng, F. Jiang, W. Chen, M. Wang, and C. Zhai, "A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data," *Energy Buildings*, vol. 216, Jun. 2020, Art. no. 109941, doi: 10.1016/j.enbuild.2020.109941.
[39] J. Ma, J. C. P. Cheng, C. Lin, Y. Tan, and J. Zhang, "Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques," *Atmos. Environ.*, vol. 214, Oct. 2019, Art. no. 116885, doi: 10.1016/j.atmosenv.2019.116885.
[40] C. Lin, L. D. Labzovskii, H. W. Leung Mak, J. C. H. Fung, A. K. H. Lau, S. T. Kenea, M. Bilal, J. D. Vande Hey, X. Lu, and J. Ma, "Observation of PM2.5 using a combination of satellite remote sensing and low-cost sensor network in Siberian urban areas with limited reference monitoring," *Atmos. Environ.*, vol. 227, Apr. 2020, Art. no. 117410, doi: 10.1016/j.atmosenv.2020.117410.
[41] G. Kundu, M. M. Islam, S. Munir, and M. F. Bari, "ACN: An associative classifier with negative rules," in *Proc. 11th IEEE Int. Conf. Comput. Sci. Eng.*, Jul. 2008, pp. 369–375, doi: 10.1109/CSE.2008.48.
[42] B. Ramasubbareddy, A. Govardhan, and A. Ramamohanreddy, "Classification Based on Positive and Negative Association Rules," *Int. J. Data Eng.*, vol. 2, no. 2, p. 84, Sep. 2011.
[43] C. Lin, A. K. H. Lau, J. C. H. Fung, Q. He, J. Ma, X. Lu, Z. Li, C. Li, R. Zuo, and A. H. S. Wong, "Decomposing the long-term variation in population exposure to outdoor PM2.5 in the greater bay area of China using satellite observations," *Remote Sens.*, vol. 11, no. 22, p. 2646, Nov. 2019, doi: 10.3390/rs11222646.
[44] J. C. P. Cheng and L. J. Ma, "A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects," *Building Environ.*, vol. 93, pp. 349–361, Nov. 2015, doi: 10.1016/j.buildenv.2015.07.019.
[45] C. Lin, A. K. H. Lau, X. Q. Lao, J. C. H. Fung, X. Lu, Z. Li, J. Ma, C. Li, and A. H. S. Wong, "A novel framework for decomposing PM2.5 variation and demographic change effects on human exposure using satellite observations," *Environ. Res.*, vol. 182, Mar. 2020, Art. no. 109120, doi: 10.1016/j.envres.2020.109120.
[46] J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan, and J. Zhang, "Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: A city management perspective," *IEEE Access*, vol. 7, pp. 148059–148072, 2019, doi: 10.1109/ACCESS.2019.2946401.
[47] J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan, "Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM," *IEEE Access*, vol. 7, pp. 107897–107907, 2019, doi: 10.1109/ACCESS.2019.2932445.

CHONG ZHAI received the master's degree from the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, in 2010. He is currently the Chief Executive Officer of Shenzhen Qianhai Bruco Consulting Company Ltd., Shenzhen, China. His research interests include smart city and artificial intelligence.

ZHENG LI received the B.Eng. degree in mechanical engineering from the Huazhong University of Science and Technology, in 2012. He was a Mechanical Engineer at Midea Group, Foshan, China. He is currently an AI Researcher at Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include machine learning, deep learning, and data mining in smart city.

FEIFENG JIANG is currently pursuing the Ph.D. degree with the Department of Architecture and Civil Engineering, City University of Hong Kong. Her main research interests include traffic safety analysis, machine learning, and big data.

JACK J. MA received the Ph.D. degree from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2016. He is currently the Chief Research Officer at the Department of Research and Development, Big Bay Innovation Research and Development Ltd., Hong Kong. His research interests include smart city, urban computing, data mining, and artificial intelligence.

ZHERUI XU received the master's degree from the School of Business, Macau University of Science and Technology, Macau, in 2014. He is currently a Senior Analyst at Shenzhen Topband Company Ltd., Shenzhen, China. His research interests include environment computing, urban computing, and artificial intelligence.