Fuzzy Systems and Data Mining V
A.J. Tallón-Ballesteros (Ed.)
IOS Press, 2019

doi:10.3233/FAIA190184

Classification-Based Data Mining Applied in Vehicle Accident Prediction

ChienHsing WU[1a], ShangWei KUO[a], Shu-Chen KAO[b]
[a] *National University of Kaohsiung, Taiwan*
[b] *Kun Shan University, Taiwan*

Abstract. The purpose of the exploratory research is to employ a classification-based data mining technique to develop a vehicle accident prediction model. Data from 2014 to 2016 was collected from the open government data of Taoyuan municipality, Taiwan, that contains five categories as the potential determinants of vehicle accident, namely temporal, environmental, human (drivers), vehicle, and miscellaneous. Each category contains various variables. The class has 11 values (e.g., head, neck, leg (foot), multiple wounds). The mining mechanism used was ID3 which is a classification-based technique. The dataset used contains 92,558 cases. Steps were conducted including data preparation, mining mechanism implementation, and validation. The results reveal that variables in human category holds the highest classification power and the environmental ones reveals the lowest. The overall prediction accuracy is 73.05%. The total number of rule discovered is 10226, of which 4088 are reliable that the conflict rate is no less than 0.5. Findings and discussions are also addressed.

Keywords: data mining, vehicle accident, prediction, open government data

Background

Taiwan has been experiencing one of the Asian countries with fast economic growth since 1980s. This results in increase of traffic volume, and is prone to the occurrence possibility of vehicle accidents. Despite of various reasons for traffic accidents, possible policies and strategies have been proposed and implemented by Taiwan government to help solve such a serious traffic and transportation problems [1, 2, 3].

Literature indicated that models used to predict accidents by adopting various data analysis techniques are proposed, such as statistic-based models [4, 5, 6] and classification-oriented models (decision tree) used to discover decision rules to detect traffic accidents [7, 8]. Variables used in the models are legal variables (e.g., alcohol law, helmet law), socioeconomic variables (e.g., number of vehicle, price of fuel) [3], vehicle variables (e.g., belt, bumpers, vehicle types) [4, 7], human variables (e.g., age, drinking, speed) [4, 8], and environment variables (e.g., traffic volume, weather, load condition, lighting) [5, 6, 8, 9]. Deepening understandings with respect to the importance and influencing ranks of variables for human, vehicle, and environment categories are particularly important to the aid of traffic policies and strategies. A prediction model discovering open government data to reveal determinants linking to vehicle accidents in Taiwan is proposed and validated in this paper.

[1]Corresponding author: Professor, Department of Information Management, 700, Kaohsiung University Rd., Nanzih District, Kaohsiung 811, Taiwan, R.O.C.; E-mail: chwu@nuk.edu.tw

1. Related Concepts

1.1 Prediction of traffic accidents

Existing models have proposed to explain the determinants that help identify and reduce possibilities of vehicle accidents. Studies are generally based on quantitative approaches, such as binary logics, regression model, Naïve Bayes model, decision tree, linear model, and general statistics. The operating categories cover four facets including human, vehicle, environment, and legal and socioeconomic, of which most studies consider the category of environment [5-9], and legal and socioeconomic from the governmental perspective. Early studies considered two or three categories whereas recent studies focused only on a singular one. For example, human, vehicle, and environment categories were considered together as the potential factors leading to traffic accidents or road safety risks in 2012, 2013, and 2015 [8]. Environment category was singularly examined for the road accident in 2017 and 2018 [5, 9]. Variables used in three categories are various depending highly upon data availability and findings are not quite consistent [6, 8]. Implications and suggestions are not consistently addressed.

1.2 Classification-based mechanism

Classification-based data mining (DM) has been successfully applied in various domain [19-21] with advantageous features of entirely a data-driven approach, learnability, high classification accuracy, and multi-context datasets [10-11]. Decision tree is a prospective mechanism when classification-based prediction model is considered [11]. The ID3 mining algorithm with a data-driven and top-down classification technique is utilized as the mining mechanism to develop decision tree and decision rules [10]. The structure of the mined decision tree is determined by the entropy of attribute (determinants), in which higher entropy implies higher power of classification.

2. Method

2.1 Research design

The three-phase process to derive research results is used (Table 1), including preparation, implementation, and validation. Preparation dealt with data collection and preprocessing which includes elimination of missing data occurred in either cases or attributes. The implementation employed the ID3 algorithm to rank entropy for attributes (or variables) while validation is to evaluate accuracy for the rules mined from the random 70% cases over the remaining 30%. The design characteristics is presented in Table 2 including 92,558 cases and all variables are discrete.

2.2 Dataset

The dataset was the open data downloaded from the Taoyuan's open government data for vehicle accident (https://data.gov.tw/dataset/45465). The variable values are written and directly entered into database by police officers. Details are presented in Table 2.

Table 1: Design characteristics

Characteristics	Description
Objective	Data mining applied in vehicle accident prediction
Datasets	Open government data of vehicle accident from 2014 to 2016
Data preprocessing	(1) Elimination of missing data (2) Confirmation of discrete data type
Mining mechanism/output	Classification-based (ID3)/Decision tree
Validation and findings	(1) Accuracy (70% for training and 30% for validation) (2) Result interpretation and discussion for vehicle accidents

Table 2: Datasets description

Dataset		(1) 2014 Jan. ~ 2016 Dec., (2) Data size = 92,558, (3) Discrete
Context: Vehicle accident	Variable category	Variables
	Temporal (2)	Year, Month
	Environmental (16)	Weather, Lighting, Road type, Road condition, Accident location, Road surface, Surface condition, Surface defect, Obstacle, Visible distance, Signal type, Signal status, Lane split facility, Sign between express lane and regular lane, Sign between express and slow lane, Sideline on road
	Human (driver) (10)	Drivers' identities, License status, Drivers' behavior, License type, Occupation, Alcohol usage, Gender, Traveling purpose, Mobile phone use (or similar devices), Hit and run
	Vehicle (3)	Vehicle type, Vehicle collision spot, Protecting equipment
	Miscellaneous (2)	Accident type and status, Main cause
	Class (Confirmed accident)	Class values (11 kinds of injuries): Head, Neck, Breast, Abdomen, Waist, Back, Hand (wrist), Leg (foot), Multiple Wounds, None, Unclear

2.3 Implementation of data mining

The ID3 algorithm [10] was utilized to produce the decision tree for the dataset. The expected information, and final gained information for a variable V is determined by the algorithm. Because of an application-oriented study, the research adopted a software tool to implement the mining algorithm, including entropy computation, decision tree illustration, decision rules listing, and prediction accuracy test.

3. Findings and Discussions

Expected entropy and ranks of attributes is presented in Table 3. The Drivers' identity (e.g., motorcycle drivers) holds the highest entropy to classify vehicle accidents, supporting the report in [8] that drivers' characteristics and behaviors are the main determinants associated with traffic accidents. The second highest determinant is Vehicle collision spot (e.g., bottom and front left end) supporting the finding that collision type leads to safety. These imply that in Taiwan the driver types and collision places are the major determinants linked to vehicle accidents with injuries. Descriptive statistics indicates that drivers of regular heavy motorcycle (41,793 cases) and collision spot on vehicle bottom (22,698 cases) are the main conditions that likely lead to conclusion with injuries of legs and heads.

For the third highest determinant (i.e., license type), the majority is small vehicle (bicycle or light motorcycle) (38,086 cases) which leads to such injuries as head and

neck; but some are unclear, or even no injury. As for the fourth highest determinant (i.e., drivers' behavior), the majority of acting behavior is moving forward and turning left. This behavior likely leads to legs or multiple wounds; but some show no injury. This finding is particularly obvious that previous studies in literature did not cover.

Figure 1 indicates that, comparatively, human category plays a more important role than other categories, implying the main cause that leads to vehicle accidents is drivers' characteristics, instead of environment. This finding signifies that simply considering the environment as the influential factor linking to traffic accidents is not sufficient. The finding is consistent with the report in [6] that speed is the only determinant of accident and that in [8] that drivers' characteristics are the main cause of traffic accident. Dataset shows that the majority of accident type is side crash (30,694 cases) and slight impact from the same side (10,477 cases). This finding supports the report in [7] that accident type is the main factor used to detect traffic accidents.

Table 3: Variable entropy and rank

Category	Variable	Entropy	Ranks	Category	Variable	Entropy	Ranks
Temporal	Month	0.0028	25	Human	Drivers' identity	0.5520	1
	Year	0.0005	33		License type	0.4882	3
Envi.	Road type	0.0161	13		Drivers' behavior	0.1308	4
	Accident location	0.0153	14		Drivers' status	0.0474	7
	Signal type	0.0061	19		Occupation	0.0405	8
	Signal status	0.0050	20		Alcohol usage	0.0314	9
	Load defect	0.0044	21		Gender	0.0180	10
	Lighting	0.0026	26		Traveling purpose	0.0173	12
	Load type	0.0065	17		Cell phone use	0.0150	15
	Lane split facility,	0.0032	23		Hit and run	0.0105	16
	Visible distance	0.0020	27	Vehicle	Vehicle type	0.0062	18
	Obstacle	0.0009	30		Collision spot	0.5387	2
	Road condition	0.0016	28		Protecting equipment	0.0179	11
	Weather	0.0011	29	Misce.	Acci. type and status	0.1291	5
	Sign (exp. and reg)	0.0037	22		Main cause	0.1050	6
	Surface condition	0.0008	31				
	Sign (exp. and slow)	0.0006	32				
	Road sideline	0.0031	24				

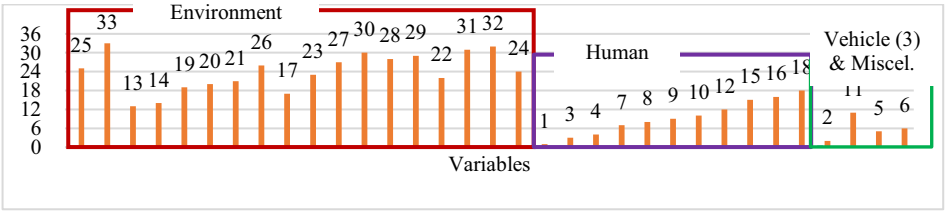

Figure 1: Classification ranks for variable categories

Part of the graphical output is shown in Figure 2. Validation results are presented in Table 4, which shows that 11,226 rules are discovered, the prediction accuracy is 73.11%, which is not unacceptable though not quite high. The rule base generated contains unreliable rules in which same condition(s) produces various conclusion. For example, a rule with the form below represents that same conditions (X21, X34, X32, X27) results in multiple conclusions (Head (3 cases), Neck (0 case), Breast (0 case), Abdomen (0 case), Waist (0 case), Back (0 case), Hand (wrist) (0 case), Leg (foot) (13 cases), Multiple

Wounds (7 cases), None (2 cases), Unclear (0 case)), where X21 is drivers' identity, X34 is accident type and status, X32 is accident cause, and X27 is occupation.

IF X21 = [C03] and X34 = [q] and X32 = [bq] and X27 = [s] THEN Y1 = [h] with majority of leg(foot) 0.5200 on (3,0,0,0,0,0,0,13,7,2,0)

Where {*X21*=Drivers' identity, *C03*=small vehicle}, {*X34*=Accident type and status, *q*=hit of vehicles}, {*X32*=Accident cause, *bq*= not noticing traffic while crossing road}, {*X27*=occupation, *s*= housewife (husband)}, {*Y1*=Main injury, *h*=leg(foot)}, 0.5200=13/(3+0+0+0+0+0+0+0+13+7+2+0)

The highest number of conclusion case is Leg (foot) (denoted by h) with 13 cases and there are 12 conflict cases that contain different conclusions. This produces a 0.5200 majority rate or reliability level. To balance the dilemma between loss of entire information if exclusion and increase of incorrect information if inclusion, the research takes on the filtering criteria that includes the rule with the number of case (or supports) if reliability level is not less than 0.5000. The result by conducting the filtering criteria reveals 4088 rules (39.98%), of which the reliability level is not less than 0.5000 and produces an average of 0.8681 reliability level.

Table 4: The mined results and validation

Items	Results (overall)	Results (adjusted by conflict data)
Dataset size	92,558 (64,791+27767)	45,690 (from 64791 cases)
Number of mined rules	10,226	4,088
Prediction accuracy	73.11%	-

Figure 2: Part of generated decision tree

Furthermore, the finding reveals that prediction accuracy is not fairly high, probably due to lots of rules that contain conflict cases. This issue may be related to the contents of original dataset. The research actually did not involve too much in the original dataset

because it is a government open data. The only thing done for the research in the pre-processing step is to eliminate the missing data. Future research may reexamine the applicability of prediction model of traffic accident with particular consideration of conflict data and the number of variables.

4. Conclusion

The exploratory research utilized the classification-based mining technique to model the vehicle accidents in open government data of Taoyuan city, Taiwan. On the one hand, the result is not quite satisfactory regarding the generated rules and prediction accuracy. It is realized that the number of attributes and their values, and the class values and its scales are quite complex for the original dataset. These will be very likely the drawbacks of the findings and validation presented in the present study as well. This issue is one of the future research focuses. On the other hand, the research considers multiple variable categories (i.e., temporal, human, environment, vehicle, and miscellaneous) as the potential determinants of vehicle accidents. Although studies considered three categories (human, environment, vehicle) [8] in early literature, two categories (human and environment [6], environment and vehicle [7]), and one category with the focus on the environment or vehicle [4, 5, 9, 8] in recent literature, our research finding discloses that human category is more important than the other two categories. It is suggested that while placing a special emphasis and effort on how to develop a safe traffic environment, vehicle drivers or users are likely the key to the reduction of traffic risks. Moreover, to enhance the applicability of classification-based mining technique, conflict data needs to be dealt with. These issues will be the future research focuses.

References

[1] Statistics of vehicle accidents, National Police Agent, Taiwan, Accessed on January, 2019, https://www.npa.gov.tw/NPAGip/wSite/ct?xItem=78478&ctNode=12878&mp=1
[2] L.B. Connelly & R. Supangan, The economic costs of road traffic crashes: Australia, states and territories, Accident Analysis and Prevention, **38** (2006), 1087-1093.
[3] G. Vorel, S.C. Kao, C.H. Wu & C.C. Wu, Determinants of traffic fatalities in Taiwan, International Journal of Information and Management Sciences, **SI-August** (2014), 233-249.
[4] J. Antona-Makoshi, K. Mikami, M. Lindkvist, J. Davidsson, & S. Schick, Accident analysis to support the development of strategies for the prevention of brain injuries in car crashes, Accident Analysis and Prevention **117** (2018) 98–105.
[5] O. Kaygisiz, M. Senbil and A. Yildiz, Influence of urban built environment on traffic accidents: The case of Eskisehir (Turkey), Case Studies on Transport Policy, **5** (2017), 306–313.
[6] V. Ratanavaraha and S. Suangka, Impacts of accident severity factors and loss values of crashes on expressways in Thailand, IATSS Research, **37** (2014), 130–136.
[7] A.C. da Cruz Figueira, C.S. Pitombo, P.T.M.S. de Oliveira & A.P.C. Larocca, Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil, Case Studies on Transport Policy, **5** (2017), 200–207.
[8] J. de Oña, R. de Oña, L. Eboli, C. Forciniti, J.L. Machado & G. Mazzulla, Analyzing the relationship among accident severity, drivers' behavior and their socio-economic characteristics in different territorial contexts, Procedia - Social and Behavioral Sciences, **160** (2014), 74–83.
[9] S.G. Charlton, N.J. Starkey & N. Malhotra, Using road markings as a continuous cue for speed choice, Accident Analysis and Prevention, **117** (2018), 288–297.
[10] J. R. Quinlan, Induction of decision tree, Machine Learning, **1** (1986), 81-106.
[11] M. Ture, F. Tokatli & I. Kurt, Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients, Expert Systems with Applications, **36(2)** (2009), 2017-2026.