TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

Survey on Severity Rate of Road Accident Assessment and Estimation using Data Mining Techniques

B.Sikander[1]
sikander.gla_mt18@gla.ac.in
Department of Computer Engineering and Applications GLA University, Mathura, India
Anant Ram[1]
anant.ram@gla.ac.in
Department of Computer Engineering and Applications GLA University, Mathura, India

Article Info
Volume 83
Page Number: 1219 - 1225
Publication Issue:
May - June 2020

Article History
Article Received*: 11August 2019*
Revised*: 18November 2019*
Accepted*: 23January 2020*
Publication*:10 May2020*

Abstract:

Traffic accidents are the world's leading source of both mortality and critical injury. A developing country such as India is often faced with a traffic accident issue. When on the road, each needs to live and protected from any event that might trigger a dead individual. To ensure the safety of life on the road, we need to perform a deep analysis of all the cause of serious and minor injury and also analyze those condition where the happen of any injury is to be prevented. we can perform classification association rules by various mining algorithms. we perform a survey regarding various mining algorithms which gives better results in various cases.

Keywords: *Association rule mining, Data mining, random tree, Road accident, Naïve Bayes.*

1. INTRODUCTION

Traffic accidents trigger about 1.25 million fatalities globally annually around 2-5 crores not major critical injuries with many related illnesses as per the World Health Organization (WHO, 2019). Almost five lakh road accidents happened in India in 2019, killing about 1.5 lakh people and injuring over five lakh citizens. According to the report given by the national crime records bureau, in 2019 almost 496,762 railways, railways crossing and road traffic collisions. these 464,674-collision records are the reason for 148,707 people killed in India by traffic. Change in the pattern can occur in the future but still, it is difficult to identify the circumstance of occurring road accidents. Furthermore, through data mining methods, we need to analyze the secret trend that affects the frequency rates of road accidents.

There are numerous classification algorithms available in the literature and also used in much research work. These algorithms are train/test type algorithms in which we build a model based on some ground valid truth and then we test such train model on some test cases through which we get our model which is able to predicate the target class.

There is numerous mining algorithm for the huge dataset which is used to find the relation between independent variables. Among different mining algorithms, the algorithm which able to find the association b/w the instance stored in the large database is association rule mining. To assess the correlation between different severe road accident factors that affect the extent of traffic accidents in India, Frequent pattern mining and predictive apriori are two popular methods in the literature.

2. RELATED WORK

Yannis et al [1], the authors used log-normal regression methods to analyze the extent of road accidents by vehicle type. The impact of this article proof that during the night, harsh weather & injuries raise the extent of the incident. The article also proof that there is a significant outcome on the category of the crash during examing the severity of the accident.

Sachin et al [2], using heterogeneous data from road accidents, the k-Modes and Latent Clustering algorithms which were used to form various homogeneous clusters. Furthermore, the Frequent Pattern growth algorithm is added to various clusters

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

developed to evaluate the method which is more effective when reducing traffic injury information heterogeneity. The study shows there are no clustering methods that are preferable to others, which implies that both clustering techniques work well in increasing the heterogeneity of incident information.

L. Mussone et al [3] perform a study in which as the road they select urban roadway intersections to evaluate the factors affecting the crash frequency. The author used two machine learning approaches which are neural network backpropagation and generalized linear mixed design. Each method proof that traffic flow is an important variable to predict the severity of the crash.

Venkata et al [4], analyze and classify accident risk factors correlated with teen driver injury frequency, a partial proportionality odds model has been created. The proposed model recommed that variables like time, age, mountainous terrain, type of suface on road & it's condition, vehicle type and access control can cause some critical damage to teen divers with a confidence level of 95%. Regarding all the variables recommed by the author which can cause of loss of teen drivers, light situation on road is also one of the important cause which can impact the injury of drivers.

Moist road conditions are strongly linked to teen driver crash frequency. Extreme weather conditions are one of the popular causes of death on the road. In such a situation navigating or controlling the vehicle is very hard. Handling a vehicle is very difficult for teen drivers.

Over speeding play a vital role in death on the road. According to the statistics, the death of teen drivers is usually because of over speeding the vehicle. Because of over speeding the vehicle, the statistics show that on highspeed highways teen drivers are usually suffer from serious injury. In accordance with the speed limits, the intersection as a base factor of injury has a negative impact on teen driver's injuries.

Debbie et al [5] Analyze the extent of the accident of Wyoming rural highway networks. Such regional networks comprise national, federal highways, municipal rural county roads, and the Wind River Indian Reserve (WRIR) road system. In accordance with Wyoming's tactical highway safety goal of reducing essential accidents (fatal and serious injury), accident occurrence was viewed as a conditional approach that categorized crashes as extreme or not significant. For each highway system, several logistic regression models have been created. Based on these observations, DOTs should implement effective policies and tailored development actions to reduce accident frequency on rural highways.

Xiaoxia et al [6] give an algorithm which is able to avoid the forward collision. in this algorithm, the author uses the fuzzy logic rule which he drives gathering the effects of driver deceleration profile. The author used the Virginia tech "100-car" dataset, from the dataset the author considers the driver's deceleration curves which show "critical event reaction braking".

Mohamed et al [7], climate impact research focused on collisions due to snow or rain. There is, however, a lack of understanding of collisions happening through haze. The author report provides a review of FS-related accidents utilizing Florida accidents dataset range 2003 to 2007. A two-stage analysis approach has been applied:

- analyze FS-related collision features w.r.t. temporary spread, important variables & forms of accident
- quantify influences on the extent of injury due to an FS-related collision.

Bismark [8], based on fixed-parameter -ve binomial & random-parameter -ve binomial models, present study performed an experimental crash rate review of the highway portion. The information used for the study were derived from Washington State's police-reported road section crash data relevant to the geographic highway, temperature, spatial features, and traffic conditions. The findings from both models were reported, analyzed and measured using a maximum of 158 highway stretches from 2008 to 2011, with 11,168 collisions.

Wei et al [9], authors used a structured probit regression model to analyze the accident frequency variables for truck drivers in the United States. Therefore, the outcomes of the analysis shows that visibility issue and poor weather enhances the possibility of higher-level injury severity in the truck driver.

Liling et al [10], authors perform statistical analysis by data mining algorithms such as k-means clustering, Naive Bayes and Apriori algorithm on the Fatal Accident dataset to analyze the correlation between fatal occurrences and other attributes such as temperature, crash, illumination, surface

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

conditions & drug drivers. The outcomes of the evaluation suggest that environmental factors including road surface, light, and temperature patterns didn't have that much impact on the fatality rate as compared to humans which are the leading cause of collision of the vehicle have a stronger impact on fatality rate.

Alireza et al [11], the impact of different causes on the frequency of SV rollover collision are examined in this report. Such considerations are correlated with driver-specific attributes, weather, automobile characteristics, road conditions, crash-specific characteristics, and traffic conditions. This research established a REGOP framework for evaluating the extent of injury caused by rollover crashes to get better info regarding the impact of these variables on the frequency of rollover collision.

Joseph et al [12] proposed a detailed six-phase speed zone regulation system, together with key aspects of each stage.

Dursun et al [13], Authors concentrated on identifying the risk factors involved with the patient, automobile, and incident that are important in creating a difference in the level of injury suffered in an automobile collision. The author identifies the correlations b/w different levels of the seriousness of injuries and the factors of risk that comes with the accident. Since introducing a rigorous sequence of in-formation fusion-based sensitivity analysis on the qualified predictive models, the researchers have noted the significance of collision-related risk factors. Sensitivity test outcomes prove that the key predictors of accident occurrence are the use of a prevention device (i.e. seat belt), the manner of collision and drug usage.

Hasan et al [14] perform research in which they identify the causes that impact the deadly two-wheeler accidents. The author used the binomial logistic regression method for his research. The results of the study indicate that the risks of rear-end are 42 times, head-on is 35 times and sideswipe crashes are 25 times more than the parameter "collision form" pedestrian.

Nima et al [15] research utilized statistical methods to analyze the affect on the extent of collision injuries of structural development variables, traffic characteristics & environmental conditions. Models of crash intensity are measured using crash data from 2007–2009 on two-lane rural highway sections in Illinois. To resolve the echelons pattern in collision dataset and also bulid a model that have direct impact on the sverity result of structural characteristics, a multilevel simuation methods is used. the author in this literature makes a hypothesis which states that the lower level includes the characteristics of the collision, on the other hand, the upper-level model included the characteristics of the segment.

Bhaven et al [16], two mutually exclusive data sets were merged and random parameters (mixed) were added to the organized logit model to be perceived as discrete diversity of the results. Analysis outcomes showed that a predictor for the extent of injuries is contained in the rain, air temperature, humidity, heat and wind speed. High severe injuries were correlated with warmer air conditions and precipitation.

TABLE 1. REVIEWS OF MINING ALGORITHMS

References	Methods	Aim	Result
[1]	LR model	It evaluate per vehicle type road accident severity	Good weather condition and crashes during night found to be the reason for increased road accident severity
[2]	LC clustering k-modes clustering FP growth technique.	It is able to recognized certain factors that causes the road accident severity.	The rules governed with respect to each cluster, do not validate the superiority over other technique.
[3]	GLMM & BPNN	Identify factors influencing the rate of collision frequency at intersections of urban roads.	Flows have a huge part to play in forecasting magnitude.

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

[4]	partial proportionality odds model	recognize the factors which basis injury brutality of teen drivers	time, age, mountainous terrain, type of suface on road & it's condition, vehicle type and access control is the factor that injury of teen drivers with a 95% confidence level.
[5]	Logistics Regression model	discover crash brutality factors on rural roadways in Wyoming	Able to find the reason which causes a crash. Factors include animal, impairment, motorcycle, mean speed and not using safety equipment like a seat belt
[6]	Fuzzy logic rules	build an algorithm to avoid the forward collision	the algorithm is given by the authors able to fulfill its aim by avoiding the collision between the vehicle and also reducing the driver inference on road while driving.
[7]	FS ordered model	Analysis of crashes cause by Obstruction of vision owing to fog and haze	An structured logit model of CV crashes might be anticipated to match up to the results of the FS ordered model.
[8]	Fixed and random parameter -ve binomial models	Analysis of high speed road such as highway, vehicle collision frequently	The outcomes in this research shows that the model which able to better in encapsulating the statistical significance of most of the parameters is random parameters -ve binomial model. The author also explains the differ ness in the dataset by comparing it with the traditional -ve binomial model
[9]	Ordered probit model	identifying various other causes through which injury severity can occur	On road visibility and worst weather condition can increase the percentage of injury severity
[10]	Apriori algorithms Naive Bayes k Means clustering	To identify variables directly related with serious accidents.	increase of fatal level is mostly caused by humans as compared to environmental factors
[11]	random-effects generalized ordered probit model	Find the cause variable for rollover crashes and observe the outcome of the proposed model	The proposed model perform well the MXL model in terms of goodness-of-fit measures.
[12]	Propose a framework for speed-zone guidelines	Design a framework for speed zone	Design a manual for each six-phase for speed zone
[13]	ANN SVM C5 LR	Find the vehicle-related and person-related factors which can cause automobile crashes	The reason which causes the prediction of injury severity is done by use of seat belt, usage of drug the way collision happened.
[14]	LR models	Determine the causes that influence motorcycle fatal crashes	As observed the factors which can cause motorcycle fatacl crash are type of collision happened, the time at which collision occur & no. of vehicle involved in collision.
[15]	Multilevel modeling techniques	propose a model which can encapsulate echelons pattern & impact of competitive variables which cause the collision of the vehicle	model confirm that hierarchical structure exist in within the data, where association exists between severity outcomes of crashes occurred on the same road segments .
[17]	LR model	Determine the correlation between the age with driver attitudes & behaviors	Related to older drivers, younger drivers found in unsafe traffic safety

3. METHODOLOGY

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

A. *Classification Algorithms*

Data analysis uses a classification algorithm to analyze the dataset through which they build a prediction models. The algorithm typed used in classification varies depending on the target variable. For this survey article, the aim factor is defined as a parameter classification with four possible outcomes (fatal, serious injury, minor injury, and non-injury) severity. The analysis problem is therefore characterized as a nominal classification problem and different data mining techniques are available based on the existing literature such random tree, Naïve Bayes, decision tree, logistics regression etc.

1. Naïve Bayes

   Popular and commonly employed supervised learning algorithms to identify information on road accidents, is the naïve Bayesian classifier. It is a statistical model that forecasts the probability for class membership based on the theorem of Bayes. Naive Bayes is an expectation approach that is utilized for categorizing and inference related to the theorem of the Bayes with the premise that each pair of variables, is independent.

2. Random Tree

   Such a tree is a collection of independent decision trees, which implies that the random tree operator functions just like the decision tree operator, but that only a random subset of attributes are required for each break. A Random Tree is a tree haggard from the collection of attainable trees at an opportunity. In this perspective, "at random" means that each tree has the same chance of being chosen from the trees set.

B. *Association Rule Mining Algorithms*

Such mining algorithm is one of the popular one which is used to find the important associations b/w the data stored in a large dataset. There is numerous data mining algorithms presented such as FP-growth, Apriori and predictive Apriori association mining algorithm is the popular and usually used algorithms in the field of analysis of accidents on road, which gives the best rules that proof the association b/w various attributes in huge database.

1. FP-Growth Algorithm

   This method is an improvement to the Apriori method. It squeezes data sets to an FP-tree, scans the server twice, fails to produce the nominee element sets in the rule mining method, and notably improves mining output. But the algorithm FP-Growth wishes to create an FP-tree containing all the datasets. This FP-tree has a strong storage space capacity.

2. Apriori Algorithm

   Apriori rule mining algorithm is the naive method of finding regular element-sets in a large database by creating a list of all conceivable object combinations and then computing support for them. The number of possible variations increases exponentially, though, as the number of items in the object collection rises, rendering this approach inefficient.

3. Predictive Apriori Algorithm

   The Apriori predictive algorithm is also used in a large database to uncover unknown and novel trends. This differs from the Apriori algorithm by integrating trust and support measures with a special factor called predictive accuracy.

C. *Data Mining Tools*

Data Mining facilitates the detection of new phenomena that are not yet known by using various data mining techniques from open source. Various platforms for data mining are currently available, such as WEKA, RAPID MINER, R, KNIME ... etc.

(a) Rapid Miner

Rapid Miner is one of Ingo Mierswa and Ralf Klinkenberg's open-source resources in data mining. Rapid Miner also knew how to apply many machine learning and data mining categorization and clustering algorithms as YALE (Another Learning Tool) based on XML.

(b) R

R is an open-source data mining tool based on C and FORTRAN programming language accepted by Ross Ihaka and Robert Gentleman for arithmetical computing and charts. R gives less support to data mining algorithms as compared to Rapid Miner and Weka, it

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

does implement a few data mining algorithms [18] [19].

(c) KNIME

KNIME is an easily operable data mining tool that contains a platform for data integration, data processing, data analysis, and exploration that runs inside IBM's Eclipse [18] [19]. KNIME is easy to extend and to add plugins.

(d) WEKA

Weka is one of the most widely used tools for finding hidden patterns established by the University of Waikato [19] [21]. Weka offers three means to use the tool: the Java API, a GUI, and a command-line interface (CLI). WEKA contains classification, clustering, association rules mining algorithms and data preprocessing tools [19] [21].

TABLE2. COMPARISON OF POPULAR DATA MINING TOOLS USED IN ROAD ACCIDENT ANALYSIS

Tools	Rapid Miner	WEKA	R	KNIME
Language	Language-Independent	Java	C, Fortran, and R	Java
Usage	Easy to use	Easiest to use	Complicated to use	Easy to use
Memory Usage	More memory	Less memory	More memory	-
Speed	Requires more memory to operate	Works faster on any machine	Works faster on any machine	-
Interface Type Supported	GUI	GUI/CLI	CLI	GUI

4. CONCLUSION

This report addresses the latest work on the assessment and prevention of road accidents. The severity of road traffic accidents continues to change over time and continually increases. The evolving and that severity of road traffic accidents lead to problems not knowing the nature of the crash, causes impacting the extent of the traffic accident, and proper management of vast volumes of data collected from different sources. Some scientists have tried to solve these problems but there are still holes in the estimation of the seriousness of road accidents and identifying contributing factors such as season and duration of the incident in which the accident occurred often. This leads to the challenges of analyzing and predicting accidents. Some of the difficulties involve modeling injuries to find suitable algorithms to predict rates of incident occurrence, planning information, processing time, and transformation. This study, therefore, identifies the appropriate algorithms, tools, review of recent studies and models on accident severity analysis and prediction in order to fill some of the gaps, which helps to extract hidden patterns of road traffic accidents in the future.

REFERENCES

[1] Yannis George, Theofilatos Athanasios, Pispiringos George, Investigation of road accident severity per vehicle type, Transportation Research Procedia 25C (2017) 2081–2088

[2] Sachin Kumar, Durga Toshniwal, Manoranjan Parida,(2016), A comparative analysis of heterogeneity in road accident data using data mining techniques, Evolving Systems.

[3] Mussone, L., Bassani, M., & Masci, P, (2017), Analysis of factors affecting the severity of crashes in urban road intersections, Accident Analysis & Prevention, 103, 112-122.

[4] Venkata R. Duddu, Venu Madhav Kukkapalli, Srinivas S. Pulugurtha, Crash risk factors associated with injury severity of teen drivers, IATSS Research 43 (2019) 37–43

[5] Debbie S. Shinstine, Shaun S. Wulff, Khaled Ksaibati, Factors associated with crash severity on rural roadways in Wyoming, journal of traffic and transportation engineering (English edition) 2016; 3(4): 308-323

[6] Xiaoxia Xiong, Meng Wang, Yingfeng Cai, Long Chen, Haneen Farah, Marjan Hagenzieker, A forward collision avoidance algorithm based on driver braking behavior, Accident Analysis and Prevention 129 (2019) 30–43

[7] Mohamed Abdel-Aty, Al-Ahad Ekram, Helai Huang, Keechoo Choi, A study on crashes related to visibility obstruction due to fog and smoke, Accident Analysis and Prevention 43 (2011) 1730– 1737

[8] Bismark R. D. K. Agbelie, A comparative empirical analysis of statistical models for evaluating highway segment crash frequency, journal of traffic and transportation engineering (English edition) 2016; 3(4): 374-379

May – June 2020
ISSN: 0193-4120 Page No. 1219 - 1225

[9] Wei Li, Wei Zhu, A dynamic simulation model of passenger flow distribution on schedule-based rail transit networks with train delays, journal of traffic and transportation engineering (English Edition) 2016; 3(4): 364-373

[10] Liling Li, Gongzhu Hu, Analysis of road traffic fatal accidents using data mining techniques, Conference Paper · June 2017 DOI: 10.1109/SERA.2017.7965753

[11] Alireza Jafari Anarkooli, Mehdi Hosseinpour, Adele Kardar, Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model, Accident Analysis and Prevention 106 (2017) 399–410

[12] K. Joseph Shrestha, Pramen P. Shrestha, Comprehensive framework for speed-zone guidelines, journal of traffic and transportation engineering (English Edition) 2016; 3(4) :352-363

[13] Dursun Delen, Leman Tomak, Kazim Topuz , Enes Eryarsoy, Investigating injury severity risk factors in auto-mobile crashes with predictive analytics and sensitivity analysis methods, Journal of Transport & Health.

[14] Hasan Mehdi Naqvi, Geetam Tiwari, Factors Contributing to Motorcycle Fatal Crashes on National Highways in India, Transportation Research Procedia 25C (2017) 2089–2102.

[15] Nima Haghighi, Xiaoyue Cathy Liua, Guohui Zhang, Richard J. Porter, Impact of roadway geometric features on crash severity on rural two-lane highways, Accident Analysis and Prevention 111 (2018) 34–42.

[16] Bhaven Naik, Li-Wei Tung, Shanshan Zhao, Aemal J. Khattak, Weather impacts on single-vehicle truck crash injury severity, Journal of Safety Research JSR-01334; No of Pages 9

[17] Alexander J Mizenko, Brian C Tefft, Lindsay S Arnold and Jurek G Grabowski, The relationship between age and driving attitudes and behaviors among older Americans, Mizenko et al. Injury Epidemiology (2015) 2:9 DOI 10.1186/s40621-015-0043-6

[18] Bhinge, A. V., (2015), A Comparative Study on Data Mining Tools (Doctoral dissertation, California State University, Sacramento).

[19] Baye Atnafu and Gagandeep Kaur, Survey on Analysis and Prediction of Road Traffic Accident Severity Levels using Data Mining Techniques in Maharashtra, India.

[20] Patel, P. S., & Desai, S. G., (2015), A Comparative Study on Data Mining Tools, International Journal of Advanced Trends in Computer Science and Engineering, 4(2).

[21] Rangra et al. (2014), comparative Study of Data Mining Tools". In Proceedings of International Conference on Advanced Research in Computer Science and Software Engineering, vol. 4, no. 6, pp. 216-223.