Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques

S. Vasavi

Abstract Road accidents may not be stopped altogether, but can be reduced. Driver emotions such as sad, happy, and anger can be one reason for accidents. At the same time, environment conditions such as weather, traffic on the road, load in the vehicle, type of road, health condition of driver, and speed can also be the reasons for accidents. Hidden patterns in accidents can be extracted so as to find the common features between accidents. This paper presents the results of the framework from the research study on road accident data of major national highways that pass through Krishna district for the year 2013 by applying machine learning techniques into analysis. These datasets collected from police stations are heterogeneous. Incomplete and erroneous values are corrected using data cleaning measures, and relevance attributes are identified using attribute selection measures. Clusters that are formed using K-medoids, and expectation maximization algorithms are then analyzed to discover hidden patterns using a priori algorithm. Results showed that the selected machine learning techniques are able to extract hidden patterns from the data. Density histograms are used for accident data visualization.

Keywords Machine learning techniques · Road accident data analysis · Preprocessing · Clustering · Association rule mining · Visualization

1 Introduction

Road safety means development and management of roads, provision of safer vehicles, and a comprehensive response to accidents [1]. Modern traffic management systems, such as real-time adjustment of traffic flow, model predictive control (MPC) technique in traffic light control, tolling strategy, etc., can be used in design and maintenance of roads, and also for producing safer vehicles. BRT system of Ahmadabad city has achieved its objective of providing a safe mode of transport

S. Vasavi (✉)
VR Siddhartha Engineering College, Kanuru, AP, India
e-mail: vasavi.movva@gmail.com

D.K. Mishra et al. (eds.), *Information and Communication Technology*,
Advances in Intelligent Systems and Computing 625,
https://doi.org/10.1007/978-981-10-5508-9_2

with more than 50% decrease in road traffic [2]. According to the National Crime Records Bureau [3], there were 39,344 road accidents, which resulted in the death of 14,966 persons. Another point of concern is that, while 8.9% of all accidents in the country occur in the state, the percentage of all deaths is higher at 10.8% t. Statistics also reveal that most accidental deaths involve people traveling in three-wheelers. More than 25% of accident deaths involving passengers of auto-rickshaws throughout the country are in Andhra Pradesh [3]. About 1,734 persons died in road accidents involving auto-rickshaws, and the state has the highest number of such deaths in the country [3]. According to the report given in [4], road accidents are the ninth leading cause of death in 2004 and expected to be fifth leading cause of death by 2030 worldwide. This paper proposes a framework that is based on the cluster analysis using K-medoids and expectation maximization (EM) and association rule mining using a priori algorithm. Association rule mining is further applied on these clusters to generate association rules. Performance is analyzed using precision and recall measures. The paper is organized as follows: Sect. 2 presents literature survey on various existing methods for accident data analysis. Methodology of proposed system is described in Sect. 3. Section 4 includes results obtained from our proposed system and analysis with respect to the performance measures. Conclusions and future work are given in Sect. 5.

2 Literature Survey

Results from the research study on applying large-scale data mining methods into analysis of traffic accidents on the Finnish roads are presented in [5]. The main intension is to show that the selected data mining methods are able to produce understandable patterns from the data, finding more fertilized information could be enhanced with more detailed datasets. The work of [6] emphasizes the importance of data mining classification algorithms in predicting the vehicle collision patterns occurred in training accident dataset. They followed a stepwise procedure which finally yields the required accident analysis results: data cleaning, data transformation, and relevance analysis. The feature selection algorithms have been explored to improve the classifier accuracy. The research work in [7] emphasizes the significance of data mining classification algorithms in predicting the factors which influence the road traffic accidents specific to injury severity. Further they applied feature selection methods to select the relevant road accident-related factors and Meta classifier Arc-X4 to improve the accuracy of the classifier. In order to improve road safety, the authors of [8] analyzed the Andalusia Complementary Road Network, by using advanced data mining techniques in order to discover hidden relationships between characteristic of the roads, ESM, and crashes. The research work in [9] is that accidents are not randomly scattered along the road network, and that drivers are not involved in accidents at random. Authors focused on the

contribution of road-related factors to accident severity in Ethiopia. Work presented in [10] is about discovering interesting rules from a set of generated rules using both association rule algorithms. Work reported in [11] is to reduce the number of road accidents in main cities of Tamil Nadu. They used WEKA tool and H-DTANN techniques in order to predict the road accident injury levels.

3 Proposed System

The main objective of this research is to investigate the role of human-, vehicle-, and infrastructure-related factors in accident severity by applying machine learning techniques on road accident data. The overall architecture of the proposed system is shown in Fig. 1. The steps include data cleaning, data transformation, relevance analysis, clustering, association rules generation, and finally performance evaluation.

3.1 Database Creation

A total of 30 attributes that focus on various criteria, such as accident-specific attributes, driver-specific attributes, FIR details, circumstance-specific attributes, and other attributes given in the FIR report, form the input dataset.

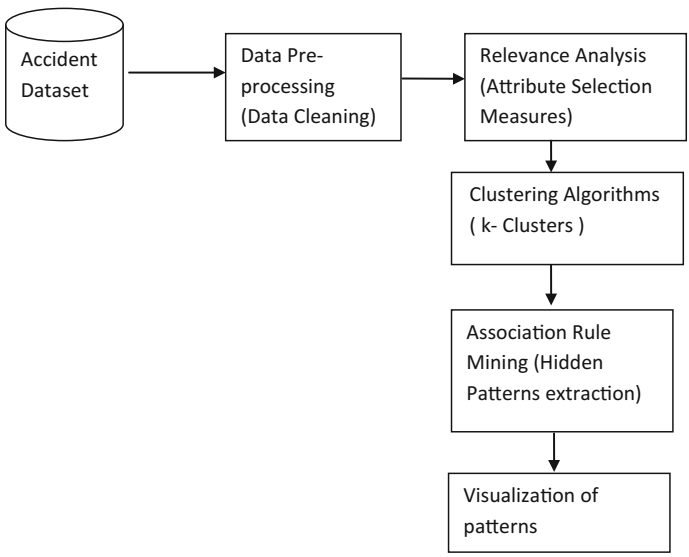

Fig. 1 Proposed system architecture

3.2 Data Preprocessing

Data preprocessing helps to remove noise, missing values, and inconsistencies. Missing values are replaced with NULL. Also each attribute data is discretized in order to make it appropriate for further analysis. Table 1 presents the data before and after transformation.

3.3 Attribute Selection Measures

Entropy measures information gain, and Gain ratio and Gini index are used to choose relevant attributes useful for performing analysis. Table 2 presents the

Table 1 Data transformation

Accident time		Accident place		Accident month		Deceased age	
Before	After	Before	After	Before	After	Before	After
8.30	Morning	Chittinagar	Chittinagar	January	January	50	Senior

Table 2 Top 20 attributes given by attribute selection measure for a Nunna dataset

Information gain	Gain ratio	Gini index	Attribute chosen
Place of accident	Any damage	Place of accident	Number injured
Any damage	Cost of damage	Any damage	Accident time
Cost of damage	Accused emotions	Cost of damage	Place of accident
Hospital reported	Place of accident	Hospital reported	Temperature
Month	Hospital reported	Month	Cost of damage
Accident type	Accident type	Accident type	Highway
Deceased emotions	No injured	Deceased emotions	Accident type
No injured	Ambulance used	No injured	Heavy traffic involved
Deceased age	Deceased emotions	Deceased age	Vehicles involved
Ambulance used	Month	Accident time	Deceased emotions
Accident time	Heavy traffic involved	Ambulance used	Accused emotions
Accused emotions	Highway	Accused emotions	Deceased age
Highway	Deceased age	Highway	Hospital reported
Vehicles involved	Vehicles involved	Weather	Month
Heavy traffic involved	Accident time	Vehicles involved	Ambulance used
Weather	Weather	Heavy traffic involved	Speed limit
Temperature	Temperature	Temperature	Road condition
Lightness	Lightness	Lightness	Lightness
Road condition	Road condition	Road condition	Weather

comparison of ranking of top 20 attributes given by each of the measure for sample dataset.

3.4 Clustering

K-medoids and expectation maximization algorithms are used for clustering, and the following clusters are formed.

Cluster 1 is the traffic cluster in which accidents happen because of low and high traffic. A total of 15% of the accidents occurred during high traffic, 76% of accidents occurred during low traffic, and 6% of accidents occurred during medium traffic.
Cluster 2 is the time of accident cluster in which accidents happen during morning, afternoon, evening, and night. A total of 32% of accidents occurred during morning time, 19.3% of accidents occurred in the afternoon, 18.5% of accidents occurred in the evening, and 29.2% of accidents occurred during nighttime.
Cluster 3 is the age of the drivers cluster in which 2.2% of accidents occurred to the age group children, 67.2% of accidents occurred to teenagers, 22.5% of accidents occurred to youth, 34.2% of accidents occurred to middle-aged people, 20.2% of accidents occurred to senior citizens, and 17% of accidents age value is missing.
Cluster 4 is the accident occurred month, in which 10.5% of accidents occurred in January, 7.7% of accidents occurred in February, 11.04% of accidents occurred in March, 9% of accidents occurred in April, 12.3% of accidents occurred in May, 11.3% of accidents occurred in June, 8.3% of accidents occurred in July, 10.24% of accidents occurred in August, 8.3% of accidents occurred in September, 7.4% of accidents occurred in October, 7.04% of accidents occurred in November, and 8.6% of accidents occurred in December.
Cluster 5 is the weather condition at the time of accident, in which 34.6% of accidents occurred when weather is cool, 33.5% of accidents occurred when weather is clear, and 31.9% of accidents occurred when weather is hot.
Cluster 6 is the lightening condition at the time of accident, in which 33% of accidents occurred when lightening is dark, 25.5% of accidents occurred in dim light, and 41.5% of accidents occurred in bright light.
Cluster 7 describes about type of accident, in which 69.5% of accidents occurred because of rash driving, 3.7% of accidents occurred because of single vehicle runoff, 0.33% of accidents occurred because of vehicle skidding, 0.33% of accidents occurred because of overlooking, 6.2% of accidents occurred because of overriding, 11.6% of accidents occurred because of hit by other vehicles, 8% of accidents occurred during lane change, and 0.34% of accidents are because of sudden turn back or animal hit, wrong direction.
Cluster 8 describes the speed limit of vehicles at the time of accident, in which 32% of accidents occurred at normal speed limit, 44.5% of accidents occurred at high speed limit, and 33.5% of accidents speed limit value is missing.

3.5 Discovery of Frequent Patterns and Association Rules

The next step is to extract hidden patterns and facts from road accident data using a priori algorithm. These hidden patterns may give analysis on various unknown risk factors behind fatal accidents and predict accident-prone areas. These rules are evaluated using support and confidence measures. Interesting measure, lift is used to rank the rules. From each of the cluster, top 20 rules are taken for analysis in this study. These rules are further visualized using density histograms.

3.6 Cluster Validation

F-measure is used for cluster analysis because it performs node-based analysis using Eqs. (1)–(3) [12].

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{1}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{2}$$

$$\text{F-Measure} = (1+\alpha)/((1/\text{Precision}) + (\alpha/\text{Recall})) \quad \text{where } \alpha = 1. \tag{3}$$

3.7 Visualization

Graphical representation techniques will help in identifying the risk of the accident immediately by government officials. Density histograms for visualizing region-wise results are generated using MATLAB software as shown in Figs. 2 and 3 for sample dataset.

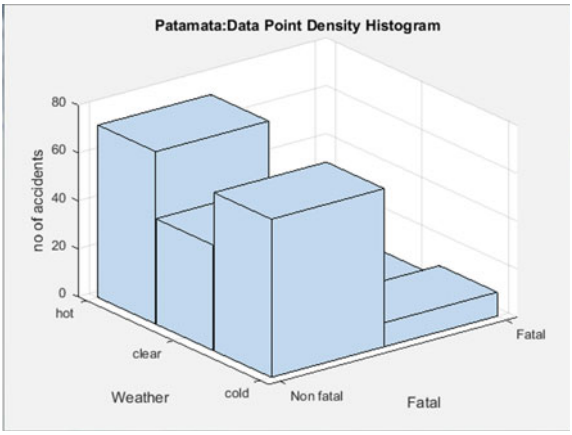

Fig. 2 Fatal versus weather

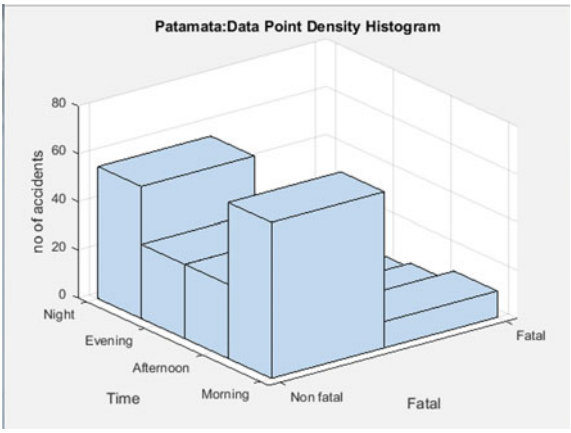

Fig. 3 Fatal versus time

Table 3 Contributing factors for accident

Contributing factor	Percentage of accidents (%)
Human–vehicle	83.64
Human	16.3
Infrastructure	0.06

Similar graphs are generated for time versus day, fatal versus month, fatal versus traffic, and fatal versus age.

4 Results and Analysis

Road accidents and injuries occur because of human fault or vehicle fault or infrastructure fault or sometimes combinations of these factors. Each of these factors individually or in combination may cause accident. It was observed from the dataset that accidents mainly occurred because of combination of human fault and vehicle fault as shown in Table 3.

Human alone factors such as "helmet and seat belt not used" are not reported in the FIRs and as such are not known. Table 4 presents top 3 contributing factors for accidents, highest being rash driving of the people.

Analysis like type of vehicles (two-wheeler, car, bus, lorry, jeep, truck, etc.) is not given in the FIR report, and as such, analysis is not done. Figures 4 and 5 present percentage distribution of accidents on various criteria, speed limit, and injury severity.

Similar analysis is done on other criteria such as distribution of accidents by time of accidents and deceased age, distribution of accidents by month and weather during the accident, distribution of accidents by lightness and speed limit, distribution of accidents by accident type (human factors), distribution of accidents by day of accident and deceased age, distribution of accidents by deceased emotions,

Table 4 Top 3 contributing factors for accidents

Contributing factor	Percentage of accidents (%)
Rash driving	62.57
Object hit	26.67
Lane change	8.1

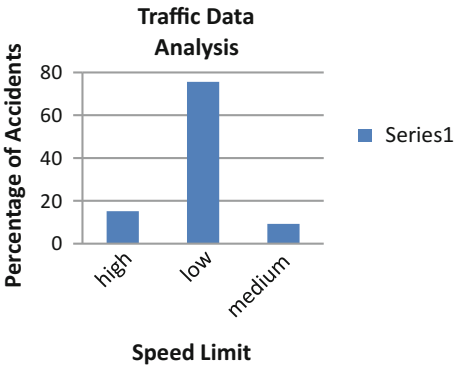

Fig. 4 Accidents by speed limit

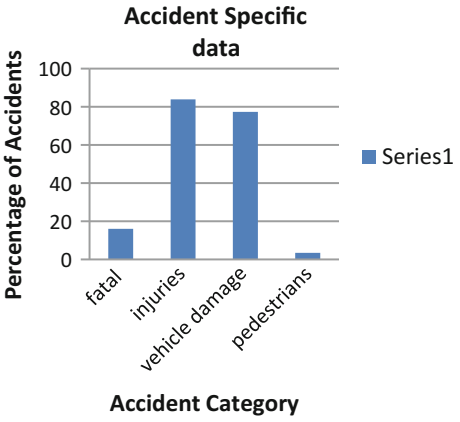

Fig. 5 Accidents by injury severity

distribution of accidents by hospital reported and ambulance used. Because of space limit, all graphs are not listed here.

K-medoids uses the cluster center to create clusters, whereas EM clustering uses the probabilities of the clusters to further calculate the optimized clusters. The results of two clusters formed from K-medoids and EM are given in Table 5.

The performance of the EM clustering is low compared to K-medoids clustering algorithm, because it uses probability measures to cluster the data. The number of iterations and runs taken to cluster the data using EM clustering is more when compared with the K-medoids clustering technique. Table 6 presents precision and recall values for both clustering algorithms.

Table 5 Comparison of clustering techniques based on emotion

K-medoids	C6 (age)	C	Y	M	S	NULL
	Obtained	1	28	35	43	11
	Expected	1	28	35	43	11
Expectation maximization	C6 (age)	C	Y	M	S	NULL
	Obtained	1	28	35	43	11
	Expected	1	28	35	43	11

Table 6 Performance measures for K-medoids and EM algorithm

Dataset	Precision		Recall		F-measure	
	EM	K-medoids	EM	K-medoids	EM	K-medoids
1504 tuples	0.5	0.8	0.4	0.6	0.45	0.69

From the data analysis, accident distribution is even in normal days, and it is observed to be higher in weekend. Accidents occurrence is high at cold nights compared to hot and clear conditions. Most accident-prone area is observed to be Kesarapalli village road and Venkateswara theater in Gannavaram. It is observed to be fatal accidents are high among the old-aged group and non-fatal in young-aged and middle-aged people. Accidents are high in the month of August and low in the month of June. Females involved in accidents are observed to be 20.16% of overall accidents to 73.45% of male.

5 Conclusions and Future Work

The aim of this paper is to generate association rules that will analyze how to discover hidden patterns that are the root causes for accidents among different combinations of attributes of a larger dataset. Density histograms for visualizing regionwise such as fatal versus weather, fatal versus time, time versus day, fatal versus month, fatal versus traffic, and fatal versus age are performed. Percentage distribution of accidents on various criteria, speed limit and injury severity, distribution of accidents by time of accidents and deceased age, distribution of accidents by month and weather during the accident, distribution of accidents by lightness and speed limit, distribution of accidents by accident type (human factors), distribution of accidents by day of accident and deceased age, distribution of accidents by deceased emotions, distribution of accidents by hospital reported and ambulance used is also made. Future work is to make analysis on road accidents' dataset by considering more features and clusters and also to use deep learning techniques so as to better cluster the records.

Acknowledgements I thank University Grants Commission (UGC), for funding this project. I also thank police authorities, Andhra Pradesh, for providing the required information. I am also thankful to the management of Siddhartha Academy for providing me resources and environment for successfully completing this project. Finally, I thank my students who helped me during the implementation of this project.

References

1. Road safety and traffic management : Report of the committee Planning Commission, Government of India in February 2007 (2007)
2. Rayle L, Pai M. Scenarios for future urbanization: carbon dioxide emissions from passenger travel in three Indian cities. Transportation Research Record: Journal of the Transportation Research Board, 2193:124–131 (2010)
3. http://www.deccanchronicle.com/130629/news-current-affairs/article/andhra-pradesh-ranked-3rd-road-accidents last accessed June 29th 2013
4. Road Accidents in India Issues & Dimensions, Ministry of Road Transport & Highways Government of India (2014)
5. SAMI AYRAMO, PASI PIRTALA, Mining road traffic accidents, Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering No. C. 2/2009 (2009)
6. S. SHANTHI, DR.R. GEETHA RAMANI, Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms, International Journal of Computer Applications (0975–8887) Volume 35– No.12, December 2011 (2011)
7. S. SHANTHI, R. GEETHA RAMANI, Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques, Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS (2012)
8. Luis Martín, Leticia Baena, Laura Garach, Griselda López, Juan de Oña Using Data Mining Techniques to Road Safety Improvement in Spanish Roads, Volume 160, Pages 607–614, XI Congreso de Ingenieria del Transporte (CIT 2014)
9. Tibebe Beshah, Shawndra Hill, Mining Road Traffic Accident Data to Improve Safety: Role of Road- related Factors on Accident Severity in Ethiopia, AAAI Spring Symposium Series (2010)
10. Amira A. El Tayeb, Vikas Pareek, Abdelaziz Araar Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-5 Issue-4, (2015)
11. K. Geetha, C. Vaishnavi Analysis on Traffic Accident Injury Level Using Classification, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 2, (2015)
12. Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques*, 2 ed, Elsevier publishers

http://www.springer.com/978-981-10-5507-2

Information and Communication Technology
Proceedings of ICICT 2016
Mishra, D.K.; Azar, A.T.; Joshi, A. (Eds.)
2018, XVIII, 340 p. 161 illus., Softcover
ISBN: 978-981-10-5507-2