Road Accidents Analysis using Data Mining

Nidhi Kalra* and Kriti Saroha**
*-**School of Information and Technology
wownidhi@gmail.com, kritisaroha@cdac.in

Abstract: Road accidents are serious threat to society. It causes both social and economic damage. Road accidents cause around 1.2 million deaths and two million accidents every year in the world. The goal of this study is to classify different regions according to various attributes like frequency of accidents and other factors. Early detection of reasons for accidents could help in taking appropriate security measures. It is analyzed that some locations are more prone to road accidents. Road accidents in those areas usually occurred due to similar features.

Keywords: Road-Accidents; Data-Mining; Clustering; Classification; K-Means; Decision-Tree; Association Rule Mining; Info- Gain Evaluator.

Introduction

Rate of accidents is increasing in India with passing years, it was recorded as 32.6% in 2013 and 36.3% in 2014. People between age group 30-45 years have been found to be more involved in accidents [3]. People killed in wars are less than people who die in road accidents. Over 1,37,000 people were killed in road accidents in 2013.In India about 1214 road crashes occur daily. 25% of accidents occur due to two wheelers, under age children has an estimation of 20 deaths daily. Uttar Pradesh has maximum rate of road rash approx. of two people die daily. Tamil Nadu has the highest rate of road crash injuries. Data mining has proven to be useful in mining road accident dataset. It finds patterns, which could infer information to evaluate rules on problems. Data mining could infer highly accident prone areas and help to propose the possible solution to the problems identified. All this would help to decrease economic and social cost of road accidents [11].

Extracted knowledge and patterns would be used to infer results and improve road security. It would be beneficial to collect maximum number of information from road accidents scene to improve analysis. Mining road accident data to analyze data and evolve patterns which could infer reasons that lead to majority of road accidents.

The main goal of studies is to classify different regions according to the reasons for road accidents on those areas. Occurrence of Road accidents is more frequent at certain locations. Early detection of these locations and factors of accidents in these locations could help stakeholders to take appropriate measures that could reduce accident rate.

Literature Survey

Data mining has been used by many studies for predicting factors affecting the road accidents. Some of them are discussed in this section.

Ref. [7] Sachin Kumar and Durga Toshniwa have used the dataset from Dehradune, Uttrakhand to perform the analysis. Duration of the dataset was from year 2009 to 2014, dataset uses 13,09,640 accident records. Authors proposed an approach of initially clustering dataset using k-means algorithm. The dataset was clustered according to frequency of accidents in different locations. Three clusters were formed with high, medium and low occurrence of accidents. Authors then used association rule mining over every cluster inferred to generate rules.

Ref. [1] Ait-Mlouk Addi, Agouti Tarik and Gharnati Fatima used dataset from Morocco, consisting of 11 attributes and 50 records. Authors introduced an approach which was divided into two modules, first is the extraction of association rules module, for extracting rules from dataset by using one of the efficacy algorithms of extraction and the second is the multi-criteria analysis approach for selecting the most relevant rules from the generated rules. Electi-tri method of multi-criteria analysis have been used to reduce the number of rules generated by Apriori association, originally 14 rules were generated out of which 12 rules were selected and 2 redundant rules were eliminated. Authors have concluded the results with accuracy of 85%.

Ref. [3] František Babic and Karin Zuskáová used dataset from UK. Authors used descriptive and predictive analysis using Decision tree and Apriori algorithms respectively. Three variations of decision tree were used and results discussed were 18.24% error with Random Forest, 14.63% error with Random Forest Big Data and 14.47% error with Gradient Boosted Classification. Descriptive analysis generated many simple and complex rules which are defined in the paper. Result shows high potential and wide scope of application for suitable data processing and analytical methods in domain of road safety.

Ref. [4] Suwarna Gothane and Dr. M. V. Sarode used dataset from India from year 2013 to 2015. Authors firstly used Info-gain evaluator that reduces the non-relevant attributes from dataset. After that, the authors used Apriori algorithm to generate

rules. Authors have discussed the results with 85% accuracy. Basic aim of this paper was to eliminate those attributes which are not related to the road accident evaluation using attribute reduction and then generate best rules to evaluate patterns and problem.

Ref. [12] RuiTian and Zhaosheng Yang and Maolei Zhang used the dataset from China of year 2005 to 2009. Authors used Rough set theory to limit the boundaries of dataset. Dataset was divided into four factors Vehicle, Road, Environment and Driver. Then Apirior algorithm was used to generate rules. Authors discussed the results with 86% of accuracy. Rough set theory works on the fuzzy boundaries, it sets limit to the boundaries and shape the data.

Ref. [13] Zhenguo Yi, Yunpeng Wang, DaxinTian, Guangquan Lu and Haiying Xia suggested an approach to mine road accidents data using neural networks. Authors considered data from 50 kilometers road which has a high crashing rate. Authors have used two data mining algorithms. Authors proposed a combinational approach of clustering and neural network. This combined algorithm is based on Hebb rule and competitive learning. The algorithm, first clusters the dataset based on seven factors. It was concluded that when neurons are between ten or fifteen, the method can give low error to evaluate the road safety.

Ref. [9] Dr. R. Geetha Ramani and S. Shanthi have suggested an approach, they considered road accidents dataset of year 2010 from Great Britain with 159417 records and 9 Attributes. The main concern of authors was to work for the pedestrians. Road safety is highly affected by the safety of pedestrians. Some of the factors of road safety that affect pedestrians are vehicle, driver, roadway and intervention variables on road accident frequency. This is complex to work on these attributes, which attracts researchers to work for pedestrian data. Authors have used Feature Selection (MIFS, Feature Ranking, CFS) to select these nine attributes from whole dataset. On pre-processed data, four variations of the decision trees: Random Tree, J48, C4.5 and Decision Stump are used to classify the dataset. Authors have discussed the results for classifiers with correctly classified instances as: Random Tree 88.9%7, C4.5 88.97%, Decision Stump 87.61%, J 48 88.77%Results also inferred that children from year 0 to 15 suffers from both fatal and minor accidents. It would be beneficial if we conduct educational campaigns for parents about road safety.

Ref. [5] Isra Al-Turaiki, Maryam Aloumi, Nour Aloumi, and Khulood Alghamdi suggested an approach for road accident mining in Saudi Arabia. Riyadh, the capital of Saudi Arabia is appeared to have maximum amount of accidents in country that is 29%. Young males form year 16 to 32 are most affected by these accidents, and hence young deaths are common in the place. According to survey from Mansuri [6], 81% of deaths are caused due to road accidents and 20% of beds in hospitals are occupied by the patients of road accidents. Dataset used is from October 2014 to October 2015. Dataset has 85,834 records for accidents, Out of which 1,808 records to injuries, 83,605 lead to no injuries and 421 records lead to deaths. Authors divided dataset into three sub dataset that are- accidents, vehicles, and parties. There are 28 attributes in accidents which describe all the details related to accidents. Vehicles dataset has 18 attributes which contains information of the vehicles involved in accidents. Parties has 14 attributes which include details of the person involved in the accidents. Authors have used three classifies on dataset they are CHAID, J48 models and Naïve Bayes. It was discussed that CHAID model starts its classification with Victim attribute, and then secondly used Vehicle Type attribute to continue splitting. It was concluded that taxi and bus drivers are less prone to accidents. J48 model also starts its decision with the Victim attribute. Results from J48 shows that Model Year attribute affect the car accident.

Results are discussed with correctly classified instances as:

CHAID 98.17%, J48 98.26%, Naive Bayes 97.06%

Authors discussed results as J48 model achieved the best performance in terms of precision.

Ref. [2] Ayushi Jain, Garima Ahuja, Anuranjana, Deepti Mehrotra suggested an approach on Indian road accidents. Authors considered data from all the union territories and states. Dataset was considered of year 2012 and has 58 attributes. Authors used two data mining algorithms. Firstly the dataset is clustered using k-means algorithm and then association rule mining is used to infer rules. Authors concluded clustering was proved to be beneficial to form clusters that determine highly prone states and territories. Clusters formed were labeled according to frequency of accidents in different areas which would be used for classification. Authors used decision tree for classification to infer best rules that affect road accidents.

Ref. [10] S.Shanthi and Dr. R. Geetha Ramani gave an approach to mine road accident dataset from India. They considered a dataset of Fatality Analysis Reporting System (FARS).They considered dataset of year 2005 to 2009. Dataset consists of 2,72,831 records and 23 attributes.

Authors have total of 63,327 records and 17 attributes after pre-processing. They divided the dataset into training and test data. Authors used the classification algorithms CART, ID3, Naive Bayes, Rnd Tree, C4.5 to classify data they are considered to be as weak evaluators.

Authors discussed their results on the basis of error rates as:

Rnd Tree 14.3% and C4.5 26.81%

Authors used Adaboost to improve the performance of classifiers.

Adaboost is a Meta learner. It is a boosting algorithm which is used to improve the accuracy of learning method. After that results were evaluated on the basis of precision, recall and ROC.

Result shows that Adaboost incorporated with Rnd improves the results and accuracy. Test data was used for evaluation. The results showed that using AdaBoost after Rnd Tree improved accuracy from 85.7% to 95.59.

Comparison Table

Table 1: describes the comparative analysis of all the papers discussed above

	A data mining approach to characterize road accident	An approach based on association rules mining to improve road safety in morocco	Descriptive and predictive mining on road accident data	Analyzing Factors, Construction of Dataset, Estimating importance of factor and generation of association rules for Indian road Accident	Method of Road Traffic Accidents Causes Analysis Based on Data Mining	A Road Safety Evaluation Method Based on Clustering Neural Network	Classifier Prediction Evaluation in Modeling Road Traffic Accident Data	Modeling Traffic Accidents in Saudi Arabia using Classification Techniques	Data Mining Approach to Analyze the Road Accidents in India	Gender Specific Classification of Road Accident Patterns through Data Mining Techniques
Algorithm	Clustering Association rule	Apriori Multiple criteria analysis	Classification Association rule	Info gain attribute evaluator Apriori algorithm	Rough set theory Association rule	Neural networks	Random Tree Decision Stump C4.5 J48	CHAID J48 Naive Bayes	CLUSTERING (K-MEANS) CLASSIFICATION (DECISION TREE)	CART ID3 Naive Bayes Rnd Tree, C4.5 Ada boost
Dataset duration	2010-2014	2002-2014	2005-2012	2013-2015	2005-2009	2007-2009	2010	2014-2015	2012	2005-2009
Accuracy	-	85%	-	85%	86%	-	87.49%	98.26%	-	85.7% to 95.59%

Proposed approach

Approach includes initially clustering dataset according to variables like frequency of accidents, number of persons involved in accidents, drunk drivers, number of lanes and weather. After that Info-gain evaluator is used over different clusters, this algorithm will reduce the attributes to be used in analysis. Algorithm removes the redundant and non-relevant attributes. Then Clustering algorithm would be performed on minimized dataset which clusters the data into similar groups. Lately Association Rule mining and Decision tree would be used on every cluster separately to generate rules on every cluster this would be beneficial to understand the problems associated with every group separately. This approach will produce the comparative analysis of both the algorithms used to make rules over dataset Proposed approach is shown below in Figure 1.
Dataset used is from US of year 2007 has total of 53 attributes and 12928 instances stances with attributes like speed limit, state code, drunk drivers included school bus included, rail included etc. All the attributes in dataset are numeric. Data set has only numerical attributes, hence we need an algorithm that is best suited for numerical data.

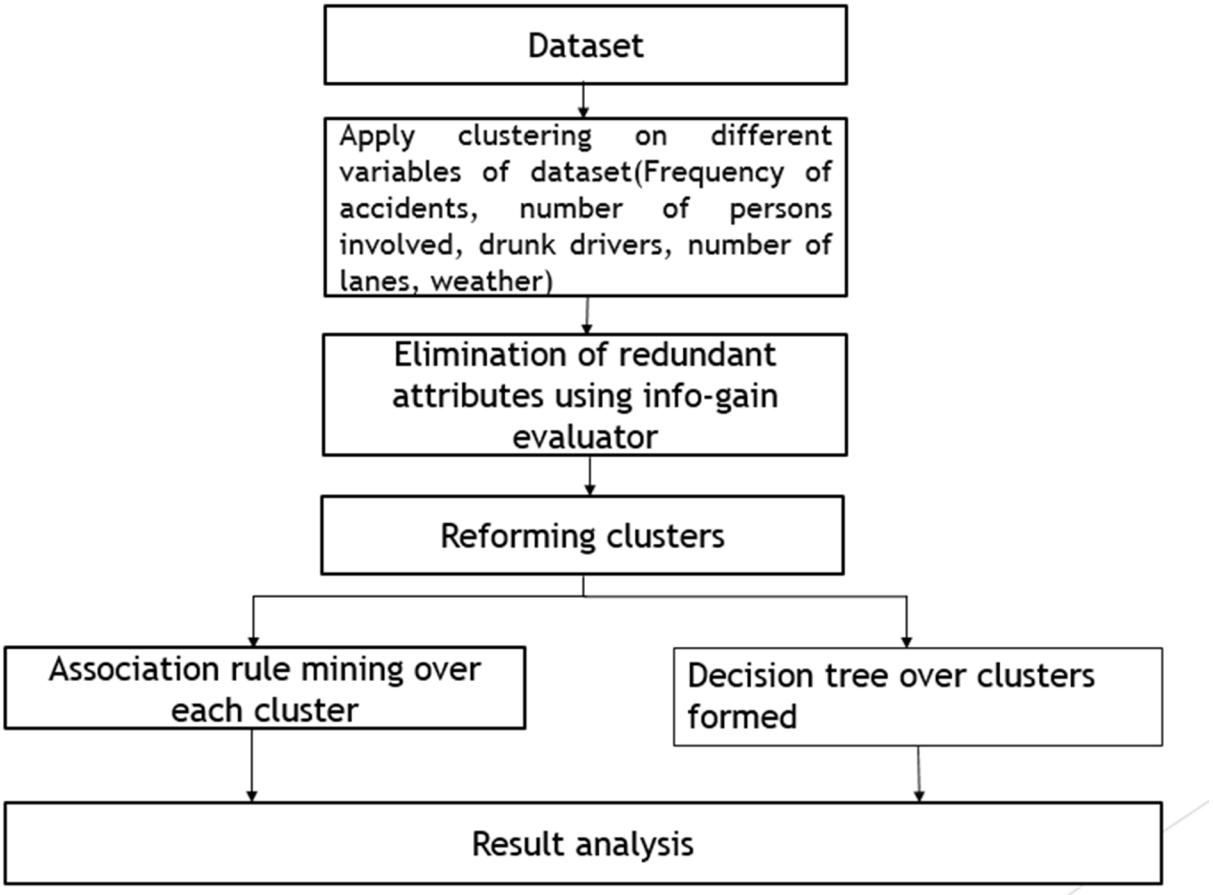

Figure 1: Proposed approach

Results

Initially dataset was clustered according to variables like frequency, number of people involved in accident, drunk drivers involved, route and weather. Three clustering algorithms were used for analysis initially. After comparisons of results, K-mean algorithm results in lowest incorrectly clustered instances and so was selected for further implementation. The comparison is shown in Table 2**.**

Info-gain Evaluator reduced the dataset according to each variable, it rank the attributes according to relevance to output variable. Dataset was reduced by removing attributes with zero ranking.

After reduction of dataset with Info-gain Evaluator, it was observed that error in clustering algorithm was reduced to some extent.

Table2: comparison of clustering algorithms

Algorithms	**Cluster instances**	**Mean square error value**	**Log likelihood**	**Incorrectly clustered instances**
K-means	7801(60%) 4010(31%) 1115(9%)			75.1044%
EM clustering	4782(37%) 3052(24%) 5092(39%)	16228.212		83.011%
Farthest first	8047(62%) 2(0%) 4877(38%)		-155.77449	79.5451%

Info-gain Evaluator reduced the dataset according to each variable, it rank the attributes according to relevance to output variable. Dataset was reduced by removing attributes with zero ranking.

After reduction of dataset with Info-gain Evaluator, it was observed that error in clustering algorithm was reduced to some extent.

Description of clusters according to different variables

Frequency of accidents: Cluster1 involves states which have medium no. of accidents in a year, Cluster2 involves states with low no. of accidents and cluster3 with high no. of accidents.

Person involved: Cluster1 includes accidents where less than ten people are involved. Cluster2 describes set of instances where exact of eleven people are involved. Cluster3 includes instances where more than eleven people are involved.

Drunken drivers: Cluster1 includes instances of less than three drunk people involved. Cluster2 groups instances of less than five people and Cluster3 for more than or equal to five people involved.

No. Of lanes: Cluster1 includes accidents which have 3,4,5,6 no. Of lanes involved in an accident, cluster2 includes accidents with 1, 2 or 7 no. Of lanes involved in accidents and cluster3 with 9 lanes involved.

Weather: Cluster1 involves accidents in rainy season and summer season, Cluster2 involves accidents in cold weather and cluster3 involves accidents in windy season.

On restructured clusters, two classification techniques were used to form rules, descriptive and predictive classification. For descriptive analysis, Decision tree was used and for predictive analysis, Apriori algorithm was used.

Description of rules by Decision tree over clusters

Frequency of accidents: Decision tree states that accidents for high prone areas occur at hilly areas and result in less than 16hrs of hospital treatment. Accidents at those areas occur after 4 in the evening. Accidents at junction take more than an hour of hospital treatment. Accidents with low light conditions and speed limit more than 45 results in fatal accidents.

Drunk drivers: Decision tree states that accidents at night occur at junction due to drunk drivers and involves approx. 2 people in accident, accidents at areas with less frequency of accidents occur mostly during morning.

No. Of Person involved: Decision tree states that accidents where more than two lanes are involved result in more than a day of treatment. Accidents occurring at more than 5 lanes occur because of drunk drivers. Accidents occurring at no junctions occur because of speed limit over 55.

No. Of lanes: Decision tree states that accidents including more than 2 lanes occur at national highways, accidents where more than 2 persons are involved occur due to drunk drivers. Accidents with more than 3 vehicles involved and more than 4 people injured involves pedestrians.

Weather: Decision tree states that in winter nights accidents occur at junction with a speed limit more than 55. Accidents in areas with less frequent accidents doesn't occur at night in months of summer, autumn or spring.

Description of rules by Apriori algorithm over clusters

Frequency of accidents: Rules concludes that most of the accidents in rainy season are not occurring at junctions, and accidents at pavements are not at junction roads.

Person involved: Rules states that, fatal accidents doesn’t involve hit-run. Accidents that include less than two people doesn’t have hit-run involved but result in fatal accidents and cluster1 doesn’t include accidents with pedestrians.

Drunken drivers: Rules conclude that most of the fatal accidents occurring rainy season. If accident involve less than three drunk people then accidents are fatal but doesn’t involve hit-run.

Number of lanes: Rules conclude that roads with no junction doesn’t cause rail and school bus accidents.

Weather: Rules conclude that accidents caused in rainy seasons are always fatal but doesn’t include hit run.

Mean-square error obtained over clusters by Decision Tree is shown in Table3.

Table 3: results obtained by decision tree

Variables	Frequency of accidents	Person involved	Drunken drivers	No. Of lanes	Weather
Accuracy achieved	90.79%	77.27%	77.28%	98%	75.90%

Confidence factor for Apriori algorithm is 0.90 and support factor is 0.85.

Conclusion

Road accidents highly affect the social and economic well-being of the world. This paper focused on finding patterns in dataset for the regions which are prone to accidents and accident features for those areas. Early detection of these reasons could help stakeholders to propose solutions to the frequent problems. This paper concluded the important of Info-gain Evaluator to find attributes which are related to accident features and different classification techniques to generate rules for

problems. It was observed that decision tree generated better rules over dataset. It was beneficial to generate clusters over every cluster separately to better understand the reasons for accidents. Analysis on different variables also results for better generation and understanding of rules.

Acknowledgment

I want to acknowledge the department and institute for their guidance and support throughout the work. I would also like to express my gratitude to the members of department for their kind assistance and cooperation for the completion of work.

References

[1] Ati-Mlouk Addi, Agouti Tarik and Gharanti Fatima “An approach based on association rules mining to improve road safety in morocco” Technology for Organization Development IEEE-2016
[2] Ayushi Jain, Garima Ahuja, Anuranjana, DeeptiMehrotra ”Data Mining Approach to Analyse the Road Accidents in India” International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) IEEE-2016
[3] Frantisek Babic and Karin Zuskacova “Descriptive and predictive mining on road accident data” International symposium on applied machine intelligence and informaticsIEEE-2016
[4] Suwarna Gothane, Dr. M. V. Sarode” Analyzing Factors, Construction of Dataset, Estimating importance of factor and generation of association rules for Indian road Accident” International Advanced Computing Conference IEEE-2016
[5] Isra Al-Turaiki, Maryam Aloumi, NourAloumi, and Khulood Alghamdi, “Modeling Traffic Accidents in Saudi Arabia using Classification Techniques” IEEE-2016
[6] F. A. Mansuri, A. H. Al-Zalabani, M. M. Zalat, and R. I. Qabshawi, “Road safety and road traffic accidents in Saudi Arabia,” Saudi Medical Journal IEEE-2015
[7] Sachin Kumar, Durga Toshniwal “A data mining approach to characterize road accident locations” Journal of Morden Transport, Springer-2015
[8] Global Status Report on Road Safety: supporting a decade of action, Geneva, World Health Organization, 2013
[9] Dr. R. GeethaRamani, S. Shanthi, “Classifier Prediction Evaluation in Modeling Road Traffic Accident Data” International Conference on Computational Intelligence and Computing Research IEEE-2012
[10] S. Shanthi and Dr. R. GeethaRamani, “Gender Specific Classification of Road Accident Patterns through Data Mining Techniques” International Conference on Advances in Engineering, Science and Management IEEE-2012
[11] Han J. and Kamber M, “Data Mining: Concepts and Techniques”, Academic Press.
[12] RuiTian, Zhaosheng Yang and Maolei Zhang “Method of Road Traffic Accidents Causes Analysis Based on Data Mining” Computational for Organizational Development IEEE-2010
[13] Zhenguo Yi1, Yunpeng Wang, DaxinTian, Guangquan Lu, Haiying Xia “A Road Safety Evaluation Method Based on Clustering Neural Network” International Conference on Optoelectronics and Image Processing IEEE-2010