International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

Traffic Risk- Safety Restraints- Awareness through Data Mining Approaches

M. Rekha Sundari, P.V.G.D. Prasad Reddy, Y. Srinivas

Abstract: ***Fatality in Road Traffic Injuries (RTI's) has been a high burden in India. Fatality rates can be affected by many factors such as types of vehicles driven, travel speeds, rates of licensure, state traffic laws, weather, and topography. Accidents can be predicted, avoided and can occur without the notice of the individual. However may be the occurrence of the accident, prevention of the fatality is at the individual risk most of the times. Surveys of RTI state that use of restraints will mostly prevent the rate of fatality in accidents. Large proportions of these RTI include Motor vehicles and mostly motor cyclists. This paper highlights the role of restraint use in reducing fatality, using Data mining approaches. Initially the personnel data is classified with two labels: Fatality and Survival using legacy classification model like Decision tree classifier. A hybrid method for classification that constructs a decision tree using Association rules is proposed. The experimental results prove that the proposed method provides better accuracy when compared to legacy methods.***

Index Terms: Association Rule Mining, Accidents, Decision tree, Fatality rate.

I. INTRODUCTION

Road traffic injuries are a major public problem and a cause of fatality around the world. Particularly in a developing country like India many people own their motor vehicles or motor cycles to carry on their day-to-day works. Hence these users make up the large proportion of fatalities in RTI's. This proportion is mostly occupied by the motor cyclists as they often share the road space with heavy vehicles like buses, trucks and rollovers and two wheelers are less visible, take less time and space for sharing a high congested traffic road. Road crash statistics report of ministry of road transport and highways 2016 state that two wheelers accounted for the highest share of road crashes in all the states in India compared to cars trucks, tempos, tractors, buses and other articulated vehicles. Two wheelers contribute to 33.8% of the road crashes where as the involvement of other vehicles is less than the stated figure. The statistics also state that 1317 crashes record 413 deaths every day or out of 55 crashes 17 deaths are recorded every hour.

The number of road crash deaths has increased by 31% from 2007 to 2017 and that of fatal road crashes have increased by 25.6% in the same period. Very recently a survey on road accidents was conducted in one of the prominent cities of Andhra Pradesh.

Manuscript published on 30 June 2019.
* Correspondence Author (s)
M. Rekha Sundari, ANITS, Visakhapatnam Andhra Pradesh, India
P.V.G.D. Prasad Reddy, Andhra University, Visakhapatnam, Andhra Pradesh, India
Y. Srinivas, GITAM University, Visakhapatnam , Andhra Pradesh, India

The statistics state that: With the total accidents of 1775, 312 are fatalities and 1562 are injuries. Of these, two wheelers occupy a major proportionate of 580 accidents 105 fatalities and 427 injuries: the highest number when compared to the other vehicles involved. While the reason for accidents may be speeding, over taking and intake of alcohol, Carelessness in the Restraint use is identified as the major cause of fatalities in the accidents. This causes death in an accident how severe or mild the accident is. Head and neck injuries are main causes of death and disability among the users of motor vehicles. Survival with head injury is typical as they incur high medical cost than any other injury. Considering this fact as the major reason technical expertise came out with a high quality helmet and high resistant seat belt to safe guard major parts of the body from injury there by reducing the fatality rate. While helmet reduces the risk of serious head injuries by reducing the impact of force or collision, seat belt protects the person in an automobile from colliding with the dash board and from flying out from the vehicle by holding the person in the seat.

Association rule mining is rule base learning methodology for discovering interesting relations between the data projected. This approach is different from other approaches like decision trees and rule based classifiers as it derives strong rules based on interesting measures like support and confidence. Innovative rules concluded from this methodology are not only applicable in market basket analyses, business analytics intrusion detection and health measures but also had its importance in all the data where relations exists between the tuples(columns).Though frequent pattern mining as gained tremendous focus from decades, our work concentrates on associative classification where associative rules are generated based on a class variable. Our study analyses huge data consisting of 84921 rows describing about the age groups, speed, restraint use, state of ejection state of injury of the person involved in the accident. The methodology implemented is known from decades but the proposed technique drives in to a new mode of conclusions.

II. RELATED WORK

Wadhwaniya et al [1] for their work conducted six rounds of road side interviews with motor cyclists involving 4872 respondents on the use of Helmet. The interviews were conducted one in two different days one in the week day and also in the weekend days on different road categories like highways, city roads, rural roads. Some respondents reported that they always wore helmets even though they are not wearing helmet at the time of interview. The objective of the work is to estimate the factors of over reporting using by bivariate and multiple logistic regression methods.

The study concluded that over reporting was least when observed unrestrained are higher and viceversa.The similar trend is observed in seat belt use. Men and youth with motorcyclists with lower levels of education were at higher risk of not wearing a helmet.

Bachani et al[2] in their study highlighted the low prevalence of helmet use in Kenya and the urgent need for efforts to improve the situation to reduce head injuries and their consequences among motor cyclists. The data was collected from the motorcyclists regarding their disinterest in wearing helmet. The data from hospitals is also collected to know the percent of motor cyclists involved in road accidents with head injuries, normal injuries and fractural injuries .The results concluded that of the most common reason highlighted by the motorcyclists for not wearing the helmet was that they are inconvenient and uncomfortable. Carlos Fernandez et a [3] implemented Apriori algorithm using Big Data architecture. Hadoop with spark distributed processing is used. Results were published comparing the traditional algorithm and the big data Association rule extraction algorithm (BDARE) using different data sets. The work concluded that the performance of the proposed algorithm in terms of time and processing capacity had great improvement when compared to traditional one.

Addi Ait-Mlouk[4] applied association rule mining on road accidents data of Morocco between 2004-2014.The dataset was obtained from the Ministry of equipment, Transport and Logistics Morocco. The concept of Multicriterion decision analysis was combined with ARM to generate hidden rules between the most common accidents. The input dataset include various factors related to the accidents such as traffic conditions, road conditions, human conditions and geographical conditions. The results concluded that the proposed approach allows decision makers to obtain rules according to their targeted class avoiding redundancy in extracted rules.

Kunyanuth et al [5] proposed a model in order to guide the students in choosing their track in the field of computer science. To select appropriate fields student registration data, course data and class learning were analyzed using data mining techniques. This research aims at developing a decision support system for guiding the students in choosing the correct field according to their abilities and interests. The data used in the experiments was collected from computer science program, Suan Sunandha Rajabhat University, during the period of 2006-2012.In the data gathering phase 4 quizzes based on computer science fields were conducted to the students in the subjects of data base, soft ware engineering ,multimedia and network and communication fields. The equal width method was used to partition the value of continuous attributes into five nominal values: VERY POOR, POOR, FAIR, GOOD and VERY GOOD. The data was analysed by using naïve Bayesian and decision trees classification techniques and the experimental results shown that naïve Bayesian is more efficient than Decision trees.

Panchal et al [6] proposed a technique to group clusters (k-means) user navigation based on the pair wise similarity measure combined with markov model. The concept of apriori algorithm is used for Web link prediction. The process predicts the Web pages to be visited by a user based on the Web pages previously visited by other users. The novel technique with clustering, markov model and apriori algorithm analyzed, evaluated and predicted future Web accesses precisely providing high accuracy.

Santra and Jayasudha [7] presented a comparison of decision tree algorithm with naïve Bayesian technique for classifying the interested users. The performance of the both the techniques is measured for web log data with session based timing, page visits, repeated user profiling, and page depth to the site length. Experimental results showed that the memory utilized is more for decision trees than naïve Bayesian and naïve Bayesian technique classified the users in a short time when compared to the decision tree.

III. METHODOLOGY

A. Dataset

The data was collected from Fatality Analysis Reporting System that contains data from 1982 to 2017 regarding traffic accidents occurred under different conditions considering the age of the person, time of the day, day of the week, region, speed, restraint use and many more factors. We have selected a small piece of data for our analysation as a startup trigger of work and this research continues in phases starting from the measures to be taken to avoid fatality to overall measures to be taken to avoid accidents or Road traffic Injuries. The dataset selected contains many attributes with different age groups of people having different age cutoffs in each group. The persons with minimum 5 years age to maximum 75 years are considered, as the work discusses about the role of restraint use of not only the driver but also the occupant of the vehicle involved in accidents. The data also consists information about the restraint use of the person involved in accident and about state of injury, whether fatal or injured.(escaped from fatality) The irrelevant attributes not needed are eliminated.

B. Preprocessing

Most of the machine learning data consists of irrelevant attributes. These attributes are not relevant in making predictions. Since the work concentrates more on the restraint use in RTI's the attributes that are irrelevant in the dataset are not considered for problem instance. The method of selecting relevant attributes in the data to model the problem is called feature selection. The attributes that are not needed for our problem statement are removed in the dataset by using correlation based feature selection technique. The correlation between each attribute and the class variable is calculated using Pearson's correlation coefficient technique and the attributes with low correlation value are dropped.

C. Association Rule Mining (ARM)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

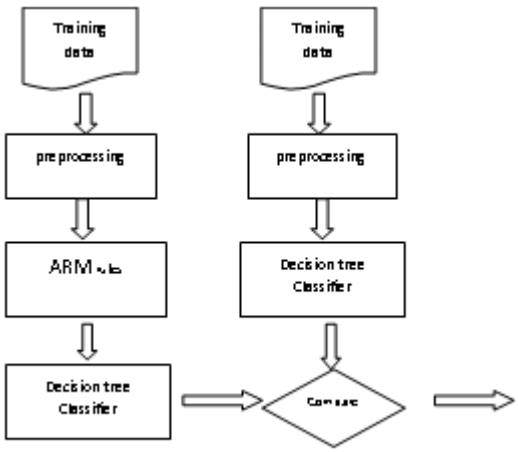

Figure 1

Association rule mining is a key pattern mining technique. The main goal of association rule mining is to solve market basket analyses problem[8] but the applications of ARM are far beyond that and have been used in various domains like medical diagnosis, bioinformatics, business analytics, web mining and further many fields. In our paper we adopt ARM in extracting rules that lead to a person fatality or survival in RTI's. This process of deriving efficient rules is performed in two ways as shown in figure 1.The Decision trees constructed on the dataset and on the ARM rules derived from the dataset are compared and the best one is summarized. PersonInjuryType is the major attribute in our dataset with two values fatal and survival from fatality. The prediction of whether a person will be dead or survived in accident deals with too many rules and it is difficult to find appropriate subset of rules to build precise and reliable predictions. In our work we concentrated on restraint use as one of the major reasons of fatality in accidents.

Association rules derive the hidden rules in a large data where the existence of antecedent determines descendant. Association rules are methodized by the help of two criterions that outline their quality: Support that ascertains the count of the behavior occurring in the rule, confidence that describes the strength of the implication.

Let P={p1,p2,p3…,pm} be a set of factors pertaining to accidents and the outcomes, W={w1,w2,w3..,wn} be the persons state of factors involved when undergone a accident and D be a collection of these factors of all pupil involved in different RTI's in different situations.

Definition 1: Support is the count, number of times an item occurred in the database.

Definition 2: Confidence gives the information about how likely Y occurred when X occurred.

The minimum support prerequisite utters the effectiveness of Association rule mining. The key motivation of using support factor shows the reality that we are interested only in rules with definite recognition. Support pertains to statistical significance while confidence pertains to strength of rules. The vast number of rules generated gives rise to the factor confidence where the rules with highest confidence are selected amongst the rest of all the applicable rules whose support values are above a definite threshold.

The dataset obtained with the Association rules generated is used to build the decision tree by calculating the attribute importance to find the best split attribute for decision making.

D. Decision tree Classifier

Input: Let D_n be a set of training dataset that reach a node n.
Step 1: If D_n contains tuples that are of same class C_n,then n is a leaf node labeled as C_n.
Step 2: If D_n contains tuples that are more than one class Place the Best split attribute of the dataset as the root of the tree.
Step 3: Recursively apply the same methodology to every subset.

The decision tree classifier is applied on the two datasets separately to estimate the accuracy of proposed method and existing method. On one side classifier is applied on the dataset generated by ARM rules, the resultant tree is named as ARMTree and on the other side classifier is applied directly on the dataset after preprocessing the resulted tree is named as DTree.

IV. IMPLEMENTATION:

A. Dataset description table

A_AGE9	Age Group 9
A_EJECT	Eject
A_PERINJ	Injury Type
A_PTYPE	Person Type
A_HELMUSE	*Motorcycle Helmet Use
A_RESTUSE	* Restraint Use

Table 1

Restraint use element 1) A_RESTUSE and 2) A_HELMUSE. A_RESTUSE focuses on belts and child seats and should be used when doing restraint use analysis on motor vehicle occupants except for motorcyclists. A_HELMUSE focuses on motorcycle helmet use.

B. Attribute name, Categories and their Descriptions

attribute A_AGE9 {1,2,3}
Age Group: <16, >16, unknown
attribute A_PTYPE {1,2 }
Person Type: driver, occupant
attribute A_RESTUSE {1,2,3}
Seat Belt Use: restrained, unrestrained, unknown
attribute A_HELMUSE {1,2,3}
Helmet Use: Helmeted, not helmeted, unknown
attribute A_EJECT {1,2,3}
Ejected or Not: Not Ejected, Ejected, unknown
attribute A_PERINJ {1,6}
Person Injury Type: fatal, survival in fatal crash

C. Generation of Association Rules

Set of rules generated by ARM with the support of maximum 100% to minimum of 1% and a minimum confidence of 75%
Size of set of large itemsets L(1): 32
Size of set of large itemsets L(2): 114

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Size of set of large itemsets L(3): 187
Size of set of large itemsets L(4): 194
Size of set of large itemsets L(5): 151
Size of set of large itemsets L(6): 89
Size of set of large itemsets L(7): 36
Size of set of large itemsets L(8): 9
Size of set of large itemsets L(9): 1

D. Sample of First Best rules Generated by ARM considering PersonInjuryType as class variable

Sno	A_AGE9	A_PTYPE	A_RESTUSE	A_EJECT		A_PERINJ	Confidence
1	A_AGE9=2	A_PTYPE=1	A_RESTUSE=2		==>	A_PERINJ=1	conf:(0.83)
3	A_AGE9=2	A_PTYPE=1	A_RESTUSE=2		==>	A_PERINJ=1	conf:(0.83)
4	A_AGE9=2	A_PTYPE=2	A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.83)
6	A_AGE9=2			A_EJECT=1	==>	A_PERINJ=6	conf:(0.82)
7		A_PTYPE=2	A_RESTUSE=1		==>	A_PERINJ=6	conf:(0.82)
8	A_AGE9=2		A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.8)
9	A_AGE9=2		A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.8)
10	A_AGE9=2	A_PTYPE=1	A_RESTUSE=2		==>	A_PERINJ=1	conf:(0.8)
11	A_AGE9=2	A_PTYPE=1	A_RESTUSE=2		==>	A_PERINJ=1	conf:(0.8)
12	A_AGE9=2		A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.8)
13	A_AGE9=2		A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.8)
14	A_AGE9=2	A_PTYPE=1	A_RESTUSE=2	A_EJECT=1	==>	A_PERINJ=1	conf:(0.8)
15	A_AGE9=2	A_PTYPE=2	A_RESTUSE=2	A_EJECT=1	==>	A_PERINJ=1	conf:(0.79)
16	A_AGE9=2		A_RESTUSE=1		==>	A_PERINJ=6	conf:(0.79)
17	A_AGE9=2		A_RESTUSE=1		==>	A_PERINJ=6	conf:(0.79)
18	A_AGE9=2	A_PTYPE=2	A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.79)
19	A_AGE9=2	A_PTYPE=2	A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.79)
20	A_AGE9=2		A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.79)
21	A_AGE9=2		A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.79)
22	A_AGE9=2		A_RESTUSE=1		==>	A_PERINJ=6	conf:(0.79)
23	A_AGE9=2		A_RESTUSE=1		==>	A_PERINJ=6	conf:(0.79)
24	A_AGE9=2	A_PTYPE=1	A_RESTUSE=1	A_EJECT=1	==>	A_PERINJ=6	conf:(0.79)
25	A_AGE9=2		A_RESTUSE=1		==>	A_PERINJ=6	conf:(0.79)

Table 2

E. Decision Tree from ARM Rules

With the data generated from the rules above we observed that the Helmet use and the Rest use can be combined to a single attribute called as restraint use, as we are considering the fatality rate on whole and not on a particular vehicle type such as car or motor bike. This combination is legal as either of the one variable is pertaining to 1 when class variable is 6 and pertaining to 2 when the class label is 1.

By implementing the algorithm stated in section-4 the best split attribute is calculated by information gain criterion. Out of the six attributes Restraint use is the attribute with highest information gain and can be named as root node. After naming the one as root node remaining all tuples are classified as the leaf nodes of the root node and further the tuple list empty. The output of the classifier is as below.

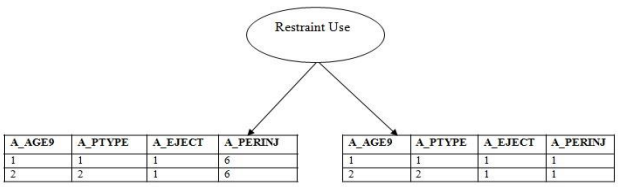

Figure 2

F. Decision tree generated without applying ARM on the dataset:

The Decision tree generated consisted of total 27 nodes with 17 terminal nodes, with a depth of 3.This huge tree is because of considering all possible facts of the data. This tree is not effective as it displays unwanted category of every attribute which needn't be considered.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

For Example the unknown category of all the attributes are not necessarily be considered for classification. This only generates meaningless data and therefore the size of the tree is huge.

V. COMPARATIVE ANALYSES

By observing the ARMtree in figure-2 and the information about the Dtree in section-4, we analyse some important advantages of constructing decision tree from ARM rules. Firstly the measure support excludes all the attributes that doesn't satisfy minimum support criterion and so the irrelevant itemsets which are less in count and doesn't contribute to meaning full information can be avoided. This is the one that doesn't happen in Dtree. This may not be applied for all the datasets such as the business data and market basket analyses data because there may be a single item in business which if sold gives high income to the head in the place of ten items bought together. The other measure that describes the strength of implied rules is confidence. This measure derives rules in the sorted order from highest to lowest value of minimum confidence set, so that the analyst can consider the rules of his interest till a particular level of satisfaction aiming towards his goal. Table 3 illustrates the confusion matrix of ARMtree with overall accuracy of 97.8 percent and Table 4 shows the confusion matrix of Dtree with 73.3 percent. This may be because of considering all unwanted categories of attributes, the constraints that are to be considered for classifying into a target class are increased numerously. By the known fact data analyzing with related data leads to fruitful results and with unrelated data leads to perturbed results, DTree has grown in an uncertain way.

Table 3

Classification Results of DTree				
Observed	Predicted			
	1	2	3	Percent Correct
1	38684	9162	98	80.7%
2	7655	14694	149	65.3%
3	3668	1902	8909	61.5%
Overall Percentag	58.9%	30.3%	10.8%	73.3%

Table 4

VI. CONCLUSION

This paper focus attention on the most commonly used data mining techniques association rules and Decision trees which allow us to distillate hidden rules from the existing datasets. The algorithms that were implemented and studied in the related work are meant for large datasets but their efficiency and accuracy are not competing when target class is specified. At this end we have implemented Decision trees from association rules generated instead of applying decision trees directly on the dataset. The results obtained show great improvement in the performance and accuracy. With the work we implemented we conclude that fatality at any point of time in RTI's can be reduced if the driver or occupant are restrained.

Regarding further research, in our future work we include large number of attributes that contribute accidents and fatality, considering various factors discussed in introduction. By analyzing the factors we conclude the elements that play a vital role in avoiding fatality in accidents.

REFERENCES

1. Wadhwaniya, Shirin, et al. "A comparison of observed and self-reported helmet use and associated factors among motorcyclists in Hyderabad city, India." Public health 144 (2017): S62-S69.
2. Bachani, A. M., et al. "Helmet wearing in Kenya: prevalence, knowledge, attitude, practice and implications." Public health 144 (2017): S23-S31.
3. Fernandez-Basso, CARLOS, M. Dolores Ruiz, And Maria J. Martin-Bautista. "Extraction of association rules using big data technologies." International Journal of Design & Nature and Ecodynamics 11.3 (2016): 178-185.
4. Ait-Mlouk, Addi, Fatima Gharnati, and Tarik Agouti. "An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety." European transport research review 9.3 (2017): e40-e40.
5. Kunyanuth. Kularbphettong, and Cholticha Tongsiri. "Mining Educational Data to Support Students' Major Selection." International Journal of Computer, Information, Systems and Control Engineering,(2014)8.1.
6. Panchal, Priyanka S., and Urmi D. Agravat. "Hybrid technique for user's web page access prediction based on Markov model." 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE, 2013.
7. Santra, A. K., and S. Jayasudha. "Classification of web log data to identify interested users using Naïve Bayesian classification." International Journal of Computer Science Issues 9.1 2012: 381-387.
8. Agarwal R. and Srikant R., "Fast algorithms for mining Association rules", VLDB'94, Chile, pp. 487–499, 1994.

Classification Results of ARMTree			
Observed	Predicted		
	1	2	Percent Correct
1	74136	1988	97.4%
2	0	8492	100.0%
Overall Percentage	80.4%	19.6%	97.8%

AUTHORS PROFILE

Dr. Rekha Sundari.M was born in in AndhraPradesh. She completed her M.Tech from GITAM University, Visakhapatnam and Ph.D from JNT University Kakinada. She is presently working as Associate Professor in Department of Information Technology and Engineering, ANITS, Visakhapatnam.Her research area includes Data Mining and Artificial Intelligence

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Prof. Prasad Reddy, P.V.G.D, was born in Andhra Pradesh. He obtained his M.Tech, in Computer Science & Technology, and Ph.D, in Computer Engineering. He is presently the BOS, CS & SE department, Andhra University, Visakhapatnam. His Research areas include Soft Computing, Software Architectures, knowledge Discovery from Databases, Image Processing, Number theory & Cryptosystems.

Prof.Srinivas Y is presently working as a Professor, in Department of Information Technology, GITAM University, Visakhapatnam. His research area includes Image Processing, Data Mining, and Software Engineering.