ANALYSIS OF ROAD TRAFFIC ACCIDENTS USING DATA MINING.

K. L. P. P. Liyanaarachchi (klppliyanaarachchi@gmail.com), E. Y. A. Charles.

Department of Computer Science, University of Jaffna.

Description of the research

Road Traffic Accidents are one of the major impact for the development of the societies around the world. Data mining algorithms can be used to analyze accident datasets to predict rules which can help to reduce road traffic accidents. The main objective of this research is to identify more accurate and useful patterns in road traffic accident data using data mining techniques. It is expected that these patterns can be utilized to take measures such that the number of accidents or the severity of the accidents can be reduced.

This research conducted a descriptive statistical analysis to understand the dataset's attributes and then used J48 decision tree classification method to construct a decision tree. To obtain useful results appropriate features were identified and in addition a measure was taken to reduce the effect of class imbalance. Decision tree constructed in this research work has shown an accuracy of 78.32%.

Dataset

Road accidents are a leading to cause for many of deaths around the world. Typically road traffic accidents are resulting in damage or injury or in fatalities. As part of this research work, accident details collected by the Traffic Police Headquarters, Colombo, Sri Lanka was obtained. These accidents occurred in Colombo district in the year 2015. A dataset with 9487 accident incidents each detailed with 55 features is used for this study. Features include a combination of numeric, ordinal and categorical data types. The collected data consists of four types of accidents, namely, Fatal, Grievous, Non-Grievous and damage only (for the vehicle). Table I. list the number of incidents in each type of accident. According to last few years of road traffic accidents data in Sri Lanka, number of fatal accidents have increased.

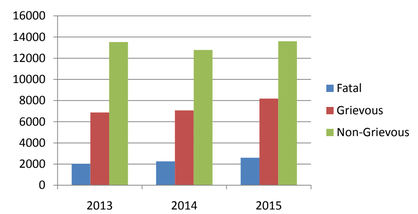


Figure I: Number of accidents in years of 2013, 2014 and 2015 in Sri Lanka respect to accident types of Fatal, Grievous, Non-Grievous and Damage only.

Table Name	Number of instances
Accident Circumstances	4672
Casual Details	1512
PNT_Accs	4672
Vehicle Details	9097

Table I: Accident database tables and number of instances

Methodology

To analyze the traffic accident dataset the following steps were performed:

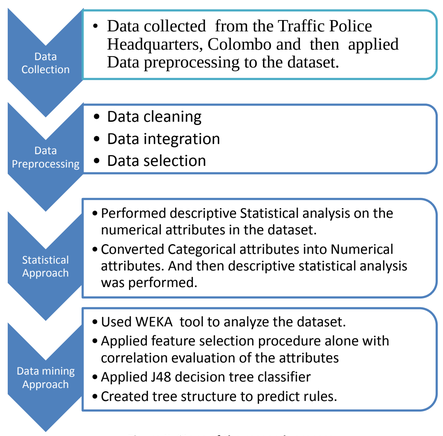


Figure II: Steps of the research process.

Results and Analysis

Statistical Approach

Descriptive statistical analysis done on some of the attributes of the dataset are given below.

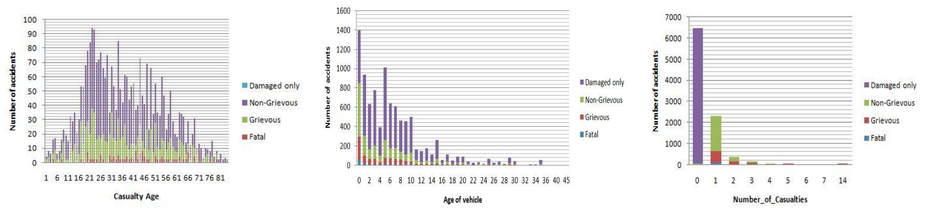


Figure III: Number of accidents with respect to the attributes of Casualty_age, Age of vehicle, Driver/Pedestrian_age and Number_Of_Casualties.

Data mining Approach

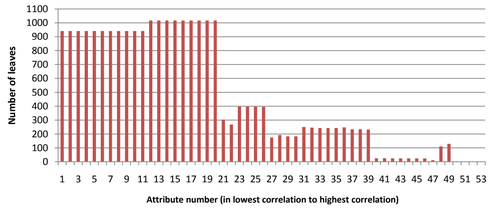
J48 Decision tree classifier was used to construct the decision tree. To obtain a useful tree, appropriate features were selected using the correlation values of each attribute.

Attribute	Correlation value	Attribute	Correlation value
Casualty Age	0.62059	Link Number	0.03915
Category	0.5807	Speed Limit Heavy Veh	0.03905
Number of Casualties	0.5423	Day of Week	0.03448
Pedestrian Location	0.45603	Vehicle Pre Crash Factor	0.0282
Protection	0.41092	Node Number	0.02796
Severity according to panel code	0.40947	Time	0.02721
Hospitalized	0.37779	Road Street Name	0.02346
Casualty Gender	0.36593	Crash Factor for Severity	0.02276
Vehicle Year of Manufacture	0.25568	Traffic Control	0.02259
License Year of Issue	0.21975	Driver/Pedestrian Age	0.02098
Validity of driving	0.19566	Driver/Pedestrian Gender	0.01975
Vehicle Ownership	0.15267	Driver Rider at Fault	0.0167
Element Type	0.14673	Nearest Lower Km Post	0.01668

Attribute	Correlation value	Attribute	Correlation value
Collision type	0.11438	Human Pre Crash Factor1	0.01656
Light Condition	0.11026	Human Pre Crash Factor2	0.01555
Other Crash Factor	0.10807	Road Number	0.01454
Work Day/Holiday	0.09893	Road Pre Crash Factor	0.01409
Number of Vehicles	0.09797	Direction of Moving	0.01352
Ped Pre Crash Factor	0.06978	Speed Limit Posted	0.01182
Number of Years Since Issue	0.06876	Location Type	0.011
Second Collision	0.06818	Distance From Node	0.01086
Speed Limit Light Veh	0.05459	Weather	0.00982
Station No	0.0542	Alcohol Test	0.00774
Age of Vehicle	0.05327	Road Surface	0.00761
East coordinate	0.04126	Urban/Rural	0.00748
North coordinate	0.04071	Distance Lower Km Post	0.00377
		DS Division	0

Table II : Correlation values of all attributes.

Feature Selection:



Preliminary results show less performance because of the class imbalance found in the dataset.

Figure IV: Number of leaves respect to the attributes that was removed.

Class Imbalance problem:

Class	Number of instances
Fatal	154
Grievous	877
Non-Grievous	2028
Damage only	6428

Class	Number of instances
Injury	3059
Damaged only	6428

Table III: Four classes converted into two classes in the target attribute (Highest severity).

The four classes were merged into two classes namely "Injury" and "Damaged only" to find the solution for the class imbalance problem. The following result was obtained after applying the feature selection on merged dataset :

Number of Leaves	218
Size of the tree	251
accuracy	78.3204
ROC Area	0.777

Table IV: Final acceptable result of the J48 classifier.

	Injury	Damaged Only
Injury	459	488
Damaged Only	129	1770

Table V: Confusion Matrix

Conclusion

J48 pruned tree

```
Vehicle_Year_of_Manufacture <=0
|   OtherCrashFactor = 2: 2 (334.0/53.0)
|   OtherCrashFactor = 0
|   |   Element_Type = 1: 2 (1.0)
|   |   Element_Type = 2: 2 (1.0)
|   |   Element_Type = 3: 2 (1.0)
|   |   Element_Type = 4: 1 (48.0/8.0)
|   |   Element_Type = 6: 1 (1.0)
|   |   Element_Type = 7: 2 (1.0)
|   |   Element_Type = 11: 2 (5.0)
|   |   Element_Type = 13: 1 (632.0)
|   |   Element_Type = 19
|   |   |   East_coordinate <= 98065: 2 (1.0)
|   |   |   East_coordinate > 98065: 1 (13.0/1.0)
|   |   Element_Type = 0
|   |   |   RoadNumber = 0: 2 (3.0)
|   |   |   RoadNumber = A002: 1 (1.0)
Vehicle_Year_of_Manufacture > 0
|   Validityofdriving = 1
|   |   Element_Type = 1
|   |   |   Second_Collision = 1
|   |   |   |   Number_of_Vehicles <= 4: 2 (153.0/33.0)
|   |   |   |   Number_of_Vehicles > 4: 1 (15.0/2.0)
|   |   |   Second_Collision = 2: 1 (11.0)
|   |   |   Second_Collision = 3: 2 (14.0/6.0)
|   |   |   Second_Collision = 0: 2 (2759.0/394.0)
|   |   Element_Type = 2
|   |   |   Second_Collision = 1: 2 (76.0/25.0)
|   |   |   Second_Collision = 2: 1 (10.0/1.0)
|   |   |   Second_Collision = 3
|   |   |   |   Licence_Year_of_Issue <= 2011: 1 (7.0)
|   |   |   |   Licence_Year_of_Issue > 2011: 2 (2.0)
|   |   |   Second_Collision = 0: 2 (1097.0/192.0)
|   |   Element_Type = 3: 2 (511.0/120.0)
|   |   Element_Type = 5: 1 (802.0/343.0)
|   |   Element_Type = 6
|   |   |   CrashFactorforSeverity = 2: 1 (1.0)
|   |   |   CrashFactorforSeverity = 4: 1 (2.0)
|   |   |   CrashFactorforSeverity = 5: 2 (4.0)
|   |   |   CrashFactorforSeverity = 6
|   |   |   |   Location_Type = 1: 1 (4.0)
|   |   |   |   Location_Type = 2: 2 (1.0)
|   |   |   CrashFactorforSeverity = 7: 1 (36.0/6.0)
|   |   |   CrashFactorforSeverity = 0: 2 (1111.0/396.0)
|   |   Element_Type = 7: 2 (73.0/18.0)
|   |   Element_Type = 8
|   |   |   DriverRideratFault = 1: 1 (80.0/37.0)
|   |   |   DriverRideratFault = 2: 2 (33.0/4.0)
|   |   Element_Type = 9
|   |   |   Number_of_Vehicles <= 1
|   |   |   |   CrashFactorforSeverity = 2: 2 (3.0)
|   |   |   |   CrashFactorforSeverity = 5: 2 (3.0)
|   |   |   |   CrashFactorforSeverity = 6: 2 (2.0)
|   |   |   |   CrashFactorforSeverity = 0: 1 (48.0/6.0)
|   |   |   Number_of_Vehicles > 1
|   |   |   |   Number_of_Vehicles <= 2: 2 (413.0/73.0)
|   |   |   |   Number_of_Vehicles > 2
|   |   |   |   |   LightCondition = 1: 2 (28.0/12.0)
|   |   |   |   |   LightCondition = 2: 1 (17.0)
|   |   |   |   |   LightCondition = 3: 1 (2.0)
|   |   |   |   |   LightCondition = 5: 2 (6.0/2.0)
|   |   Element_Type = 10
|   |   |   Vehicle_Year_of_Manufacture <= 2006: 1 (4.0)
|   |   |   Vehicle_Year_of_Manufacture > 2006: 2 (3.0)
|   |   Element_Type = 11
|   |   |   LightCondition = 1: 2 (8.0)
|   |   |   LightCondition = 5: 1 (1.0)
|   Validityofdriving = 2
|   |   Element_Type = 1
|   |   |   Traffic_Control = 1: 2 (6.0)
|   |   |   Traffic_Control = 2: 2 (12.0)
|   |   |   Traffic_Control = 3: 1 (4.0)
|   |   |   Traffic_Control = 4: 1 (2.0)
|   |   |   Traffic_Control = 6: 2 (77.0/18.0)
|   |   Element_Type = 2
|   |   |   LightCondition = 1: 2 (21.0/4.0)
|   |   |   LightCondition = 2: 2 (1.0)
|   |   |   LightCondition = 3: 2 (2.0)
|   |   |   LightCondition = 4: 1 (1.0)
|   |   |   LightCondition = 5
|   |   |   |   VehicleOwnership = 1
|   |   |   |   |   Number_of_Vehicles <= 1: 2 (1.0)
|   |   |   |   |   Number_of_Vehicles > 1: 1 (11.0)
|   |   |   |   VehicleOwnership = 3: 2 (1.0)
|   |   Element_Type = 3
|   |   |   Direction_of_Moving = N
|   |   |   |   Vehicle_Year_of_Manufacture <= 2010: 1 (4.0)
|   |   |   |   Vehicle_Year_of_Manufacture > 2010: 2 (1.0)
|   |   |   Direction_of_Moving = NE: 1 (5.0)
|   |   |   Direction_of_Moving = W: 2 (2.0)
|   |   |   Direction_of_Moving = E: 2 (5.0)
|   |   |   Direction_of_Moving = S
|   |   |   |   RoadNumber = 0: 2 (8.0/1.0)
|   |   |   |   RoadNumber = A205: 2 (1.0)
|   |   |   |   RoadNumber = A224: 1 (2.0)
|   |   |   Direction_of_Moving = SW: 1 (1.0)
|   |   Element_Type = 5: 1 (137.0/16.0)
|   |   Element_Type = 6: 1 (130.0/42.0)
|   |   Element_Type = 7
|   |   |   Day_of_Week = 3: 2 (2.0)
|   |   |   Day_of_Week = 6: 1 (1.0)
|   |   Element_Type = 8
|   |   |   East_coordinate <= 100170: 2 (2.0)
|   |   |   East_coordinate > 100170: 1 (1.0)
|   |   Element_Type = 9
|   |   |   Location_Type = 1: 2 (12.0/3.0)
|   |   |   Location_Type = 2: 1 (15.0)
|   |   |   Location_Type = 3: 2 (4.0)
|   |   |   Location_Type = 5: 1 (1.0)
|   |   |   Location_Type = 7: 2 (2.0)
|   Validityofdriving = 3: 2 (31.0/13.0)
|   Validityofdriving = 4: 2 (216.0/66.0)
|   Validityofdriving = 5
|   |   WorkDay/Holiday = 1: 2 (2.0)
|   |   WorkDay/Holiday = 2: 1 (1.0)
|   Validityofdriving = 0
|   |   Second_Collision = 1: 2 (1.0)
|   |   Second_Collision = 2: 1 (2.0)
|   |   Second_Collision = 3: 1 (4.0)
|   |   Second_Collision = 0: 2 (374.0/31.0)
```

This research study mainly considered in finding more accurate and useful patterns on the Road Traffic Accident data by using J48 decision tree classifier. The classifier showed good results on the dataset and found some useful patterns.

Example rules:

- IF "**Validity of Driving**" = 1 *(Valid license for the vehicle)* AND "**Element Type**" = 9 (*Private bus*) AND "**Number of vehicles**" > 2 AND "**Light Condition**" = 2 *(Night, no street lighting)*. THEN "**Highest Severity**" = 1 *(Injure).*
  - *correctly classified number of instances = 17/459*
- IF AND "**Validity of Driving**" = 1 *(Valid license for the vehicle)* AND "**Element Type**" = 8 *(SLTB bus)* AND "**Driver Rider At Fault**" = 1 *(Yes).* THEN "**Highest Severity**" = 1 *(Injure).*
  - *correctly classified number of instances = 80/459*
- IF AND "**Validity of driving**" = 2 *(Without valid license for the vehicle)* AND "**Element Type**" = 2 *(Dual purpose vehicle)* AND "**Light Condition**" = 5 *(Night, good street lighting)* AND "**Vehicle Ownership**" = 3 *(Government vehicle).* THEN "Highest Severity" = 2 *(Damage Only)* .
  - *correctly classified number of instances = 1/ 1770*
- IF AND "**Validity of Driving**" = 2 *(Without valid license for the vehicle)* AND "**Element Type**" = 9 *(Private bus)* AND "**Location Type**" = 2 *(4-leg junction).* THEN "Highest Severity" = 1 *(Injure).*
  - *correctly classified number of instances = 15/459*

This research work can be extended to analyze the dataset further based on the obtained results. The rules obtained can be used by traffic department of Police, to prevent accidents or to reduce the severity of injuries and also to develop automated traffic control applications.