



Analysis on Traffic Accident Injury Level Using Classification

¹K. Geetha, ²C. Vaishnavi

¹Assistant Professor (SG), ²MPhil Research Scholar
Department of Information Technology, Dr. N.G.P. Arts and Science College
Coimbatore, Tamil Nadu, India

Abstract: *Traffic Accidents are occurring due to development of automobile industry and the accidents are unavoidable even the traffic rules are very strictly maintained. Data mining algorithm is applied to model the traffic accident injury level by using traffic accident dataset. It helped by obtaining the characteristics of drivers behavior, road condition and weather condition, Accident severity that are connected with different injury severities and death. This paper presents some models to predict the severity of injury using some data mining algorithms. The study focused on collecting the real data from previous research and obtains the injury severity level of traffic accident data.*

Keywords used: *Traffic Accidents, Injuries, PART in WEKA, Classifiers, Hybrid Decision Tree Artificial Neural Network.*

I. INTRODUCTION

1.1 Data Mining

Generally, data mining (also called knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1.2 How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction. Different levels of analysis are available.

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k \geq 1$). Sometimes called the k -nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

II. PROBLEM STATEMENT

The purpose of this investigation and analysis is to reduce the number of road accidents in main city of tamilnadu by finding risks and circumstances which can be shown to be regular contributing factors to road accidents. The most common factors and circumstances related to road accidents can be established to obtain a large dataset which records, in detail, every reported road accident which resulted shows injury severity in the main city of tamilnadu.

The information is either text or numerical formats. However the information in the dataset is relatively unsorted and is therefore in an unmanageable format, meaning no conclusions can be gained regarding road accidents. We planned to get all road accident data and use data mining WEKA tools and H-DTANN techniques to predict the road accident injury levels.

2.1 Related Work

It is estimated that the annual cost to the NHS due to road traffic accidents is over £1billion per year .Tremendous amount of money goes to medical area due to these accidents. On a larger scale the World Health Organization, in conjunction with the World Bank, concluded from their analysis that over 1million people die worldwide each year as a result of road traffic crashes and collisions and injuries that by 2020 road traffic accidents could overtake HIV and Tuberculosis to rank third in the causes of premature death and disability around the world.

[1] To identify statistically significant factors using a logistic regression model that predict the probabilities of crashes and injury crashes aiming at using these models it perform a risk assessment of a given region. This model describes a site by its land use activity, road side design, use of traffic control devices and traffic exposure. Their studies focused on village sites are less hazardous than residential and shopping sites in city.

[2] Classification and regression tree (CART) and negative binomial regression models to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. This study focused on Automobile industry to help to improve vehicle safety and some environmental organizations are help to reduce the pollution to minimize the traffic accidents and traffic congestion.

[3] For two RTA severity categories, various algorithm used for improve the individual classifier. Using neural network and decision tree individual classifiers, three different approaches were applied: classifier fusion based on the Dempster–Shafer algorithm, the Bayesian procedure, and logistic model; data ensemble fusion based on arcing and bagging; and clustering based on the *k*-means algorithm. Their empirical results show that a clustering based classification algorithm works optimal for road traffic accident classification in Korea.

[4] The statistical properties of four regression models: two conventional linear regression models and two Poisson regression models in terms of their ability to model vehicle accidents and highway geometric design relationships. Highway Safety Information System (HSIS) has Roadway and truck accident data that have been employed to illustrate the use and the limitations of these models. The Poisson regression models, on the other hand, possess most of the desirable statistical properties in developing the relationships.

[5] A historical RTA data, including 4,658 accident records at the Addis Ababa Traffic Office, the records used to investigate the analysis of accident severity in Addis Ababa, Ethiopia. Using the DT technique and applying the Knowledge SEEKER algorithm of the Knowledge STUDIO data mining tool, the developed model classified the accident data based on accident severity & classified into four classes: fatal injury, serious injury, slight injury, and property damage.

[6] To analyze accident data Non-parametric Classification tree techniques is used from the year 2001 for Taipei, Taiwan. A CART model was developed to establish the relationship between injury severity and driver/vehicle characteristics, Highway/environment variables and accident variables. The most important variable associated with crash severity was the vehicle type, with pedestrians, motor cycles, and bicyclists having the highest injury risks of all driver types of accidents in the RTAs.

[7] A data mining study to classify driver responsibility levels in traffic accidents in Addis Ababa Traffic office. The study focused on identifying the important factors influencing the level of driver responsibility, and used the RTA dataset of the Addis Ababa Traffic Control and Investigation Department (AATCID).The WEKA data mining tool was used to build the decision tree (using the ID3 and J48 algorithms) and MLP (back propagation algorithm) predictive models are used here.

[8] A combination of cluster analysis, regression analysis, and geographical information system (GIS) techniques to group homogeneous accident data and also it estimates the number of traffic accidents, and assess Road Traffic Accident risk in Hong Kong.

III. METHODOLOGY

The set of queries relevant to this study was analyzed and gathered based on previous research. These included queries such as what percentage of accidents resulted in a fatality, how many accidents involved children, adults etc. When travelling greater than 50 mph, what kind of accidents occurs, which gender and age group is more likely to have an accident and which four attributes are contributed to accidents involving more than three vehicles.

The initial point of impact has 9 categories: no damage/non-collision, front, right side, left side, back, front right corner, front left corner, back right corner, back left corner. The remaining input variables are: drivers' age, gender, alcohol usages, restraint system, eject vehicle body type, vehicle role, vehicle age, rollover, road surface condition, light condition, whether condition, injuries details etc.

The research used the WEKA version 3-5-8 tool to build the decision tree (using the J48 algorithm) and rule induction (using PART algorithm) techniques. For the hybrid decision tree–ANN, we used the same hybrid learning algorithms and parameters setting as we used for Artificial Neural Network (except for the number of hidden neurons). Experiments were performed better with different number of hidden neurons and models were selected with the highest classification accuracy for the output class.

Table 1: Relevant categorical attributes description

Attribute Name	Description
DriverAge	The age of the driver
DriverExperience	The experience of driving

VehicleType	The Accident Vehicle Type(ex:two wheeler)
Subcity	The accident occurred sub city
Particulararea	The accident occurred area(ex: college)
Roadseparation	Road Segments separation
Roadorientation	How the road is oriented
RoadJunction	Type of road junction
Roadsurfacetype	Whether road surface is asphalt or ground
Roadsurfcondition	Try or muddy or wet
Weathercondition	Any of weather type
Lightcondition	The light condition
Accidentseverity	The severity of the accident

IV. EXPERIMENTAL EVALUATION

To predict accident severity, various classification models were built using decision tree (DT), naive Bayes, and K-nearest neighbor classifiers. They automatically handle interactions between variables and identify important variables. After finding important variables the assessment of the data and selecting the predictive models to be used, a series of experiments were performed in it.

Extensive data pre-processing resulted in a clean dataset containing 18,288 accidents with no missing values. The class label ('AccidentSeverity') had four nominal values which are: 'Fatal,' 'Severeinjury,' 'Slightinjury,' or 'Propertyloss.' During data exploration, different numbers of attributes were selected by different feature selection techniques.

Since WEKA's explorer generally chooses reasonable defaults, the J48 decision tree algorithm was performed using its default parameters: a confidence interval of 0.25, pruning allowed, and a minimum number of objects for a leaf of 3. Using ten-fold cross-validation training and testing also done. In the first experiment, the 18,288-accident dataset with 13 attributes, including 12 independent variables and one dependent variable (the class-label attribute 'AccidentSeverity'), were fed to WEKA's explorer. The J48 classifier was used and an accuracy of 80.641 was achieved.

In the second and third experiments, the same input, instances, and attributes were fed to WEKA Explorer. Using the naive Bayes classifier, an accuracy of 79.867% was achieved. Using the K-nearest neighbor's classifier (IBK), an accuracy of 81.231% was achieved.

For Machine Learning Paradigms the input and output variables are considered for building the model. There are no conflicts between the any attributes since each variable has their own characteristics. The variables are already categorized and are represented by numbers. The manner in which the collision occurred has 7 categories: not collision, rear end, head on, rear to rear, angle, sideswipe same direction, and sideswipe opposite direction.

For these 7 categories the distribution of the fatal injury is as follows: 0.56% for not collision, 0.08% for rear-end collision, 1.54% for head-on collision, 0.00% for rear-to-rear collision, 0.20% for angle collision, 0.08% for sideswipe same direction collision, 0.49% for sideswipe opposite direction collision. Since the highest percent of fatal injury is on head-on collision; therefore, the dataset was narrowed down to head-on collision only. Head-on collision has a total number of 10,386 records. There are 160 records of head-on collision with fatal injury; all of these 160 records have the initial point of impact categorized as front. Vehicle age with values 37, 41, 46 and 56 each has only one record and these were the only records representing such old cars.

These four records were therefore deleted from the dataset since they were clear outliers. Thus, finally, the dataset for modeling had 10,247 records. There were 5,171 (50.46%) records with no injury, 2138 (20.86%) records with possible injury, 1721 (16.80%) records with non-incapacitating injury, 1057 (10.32%) records with incapacitating injury, and 160 (1.56%) records with fatal injury.

We have separated each output class and used one against-all approach. This approach selects the positive class as one output class, and all the other classes' combination to be the negative class. We set the output value of the positive class to 1, and the (combined) negative class (es) to 0. We divided the datasets randomly into 60%, 20%, and 20% for training, cross validation, and testing respectively.

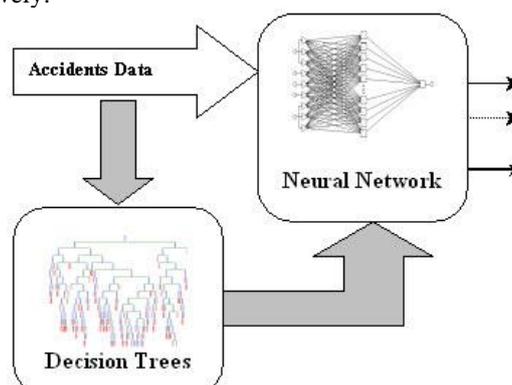


Fig 1: Hybrid concurrent decision tree-Artificial Neural Network model for accident data

V. RESULTS

From this study the results shows compare to other algorithms, classifiers in WEKA and Hybrid Decision Tree Artificial Neural Network are performed well. Decision Tree (J48), Navies Bayes-Nearest Neighbors are Performed by WEKA Tool. In Decision Tree, Naive Bayes-nearest Neighbors accuracy is 80.641%, 79.867%, 81.231%.

The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test. An entirely random test (i.e., no better at identifying true positives than flipping a coin) has an AUC of 0.5, while a perfect test (i.e., one with zero false positives or negatives) has an AUC of 1.00.

In all those classifiers (Decision Tree (J48), naive Bayes-Nearest Neighbor) AUC's performed with Property Loss, Slight injury, Fatal injury, Severe injury. Comparing the performance of classifiers K-Nearest Neighbor performs well. For fatal injury and severe injury it reaches 0.959, 0.921 so it comes close to 1.

In all cases, the AUCs were significantly > 0.5. These results indicate that all models predicted new instances well. PART is a class for generating a decision list in WEKA. Knowledge/pattern identification is done by using PART Algorithm. To identify significant rules, PART was run on the Traffic accident dataset with different numbers of attributes. Ten-fold cross validation was used for testing and the minimum number of objects in a leaf was set to twenty. The accuracy of the algorithm in generating the rules was 79.942.

For Ex:

PART decision list

Road Orientation = Straight Plain AND Road
Junction = Roundabout AND
Sub city = Kolfe: Fatal (4.0/0.0)
Road Orientation = Straight Plain AND
Sub city = Arada AND
Road Separation = Bidirectional Property Loss (52.19/18.01)

Hybrid- Decision Tree Artificial Neural Network (H-DTANN)

From the experiment results, for no injury class the best model had 72 hidden neurons, with training and testing performance of 82.95% and 63.49% respectively. For possible injury class, the best model had 95 hidden neurons with training and testing performance of 73.89% and 69.10% respectively. For non-incapacitating injury class, the best model had 109 hidden neurons with training and testing performance of 70.68% and 61.78% respectively.

For incapacitating injury class, the best model had 102 hidden neurons, with training and testing performance of 74.92% and 75.59% respectively. For fatal injury class, the best model had 76 hidden neurons with training and testing performance of 92.43% and 90.00% respectively. These are the best models out of multiple experiments. For non-incapacitating injury, incapacitating injury, and fatal injury classes, the hybrid DTANN outperformed compared to both ANN and DT. Hybrid Decision Tree artificial neural network also reached accuracy high in fatal injury.

The WEKA tool and hybrid decision tree algorithm both performed with good accuracy level for predicting the injury level.

VI. CONCLUSION

A thorough literature review deals with relationship between road characteristics and RTA severity in main city of tamilnadu. The RTA is eager to continue the study to identify areas of interest that should be given resources for traffic safety. Finally, knowledge Representation was presented in the form of rules using the PART algorithm of WEKA. Hybrid decision tree - neural network for predicting 'injury severity' in head-on front impact point collisions. The classification accuracy on the test results reveals that, for non-incapacitating injury, incapacitating injury, and fatal injury classes, the hybrid approach performed better than neural network, decision trees and support vector machines.

REFERENCES

- [1] Ossenbruggen, P. J., J. Pendharkar, et al. (2001), "Roadway safety in rural and small urbanized areas", Accidents Analysis and Prevention 33(4): 485-498.
- [2] Chang, L. and W. Chen (2005), "Data mining of tree-based models to analyze freeway accident frequency", Journal of Safety Research 36: 365-375.
- [3] Sohn, S. and S. Lee (2002), "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea", Safety Science 41(1): 1-14.
- [4] Miaou, S.P. and Harry, L. (1993), "Modeling vehicle accidents and highway geometric design relationships", Accidents Analysis and Prevention, 25 (6), pp. 689-709.
- [5] Tibebe B. Tesema, Abraham A. And Grosan C (2005) "Rule Mining and Classification of Road Traffic Accidents Using Adaptive Regression Trees", International Journal of Simulation Vol. 6 No 10 and 11, 2005.
- [6] Chang, L. and H. Wang (2006), "Analysis of traffic injury severity: An application of non-parametric classification tree techniques Accident analysis and prevention ", Accident analysis and prevention 38(5): 1019-1027.
- [7] Zelalem, R. (2009), "Determining the degree of driver's responsibility for car accident: the case of Addis Ababa traffic office", Addis Ababa, Addis Ababa University.
- [8] Ng, K. S., W. T. Hung, et al. (2002), "An algorithm for assessing the risk of traffic accidents", Journal of Safety Research 33: 387-410.