Accident Severity Prediction Using Data Mining Methods

Article *in* International Journal of Scientific Research in Computer Science Engineering and Information Technology · March 2019

DOI: 10.32628/CSEIT195293

5 authors, including:

Gaurav Gandhi

150 PUBLICATIONS **958** CITATIONS

SEE PROFILE

International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195293

Accident Severity Prediction Using Data Mining Methods

S. Ramya[1], SK. Reshma[2] , V. Dhatri Manogna[3], Y. Satya Saroja[4], Dr. G. Sanjay Gandhi[5]

[1,2,3,4]B.Tech Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Namburu, Andhra Pradesh, India

[5]Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India

ABSTRACT

The smart city concept provides opportunities to handle urban problems, and also to improve the citizens' living environment. In recent years, road traffic accidents (RTAs) have become one of the largest national health issues in the world and it is leading cause for deaths. The burden of road accident casualties and damage is much higher in developing countries than in developed countries. Many factors (driver, environment, vehicle, etc.) are related to traffic accidents, some of those factors are more important in determining the accident severity than others. The analytical data mining solutions can significantly be employed to determine and predict such influential factors among human, vehicle and environmental factors. In this research, the classification technique i.e., Random forest algorithm is used to identify relevant patterns and for classifying the type of accident severity of various traffic accidents with the help of influential environmental features of RTAs that can be used to build the prediction model. This technique was tested using a real dataset. A decision system has been built using the model generated by the Random Forest technique that will help decision makers to enhance the decision making process by predicting the severity of the accident.

Keywords : Decision Making, Accidents Severity Prediction, Random Forest Algorithm.

I. INTRODUCTION

Traffic accidents cause serious threat to the human life worldwide. In order to take necessary actions to control this ever-growing problem extensive research has been carried out into the prediction of traffic accidents in both developed and developing countries using various statistical techniques. Realizing traffic accidents as a preventable problem developed countries have implemented different policies and measures to reduce this problem. These include enforcement, education, training and engineering improvements.

Unlike developed countries, the problem of traffic accidents in developing countries is still considered as a matter of fate or unavoidable cost of development. Without remarkable efforts to enhance traffic safety in developing countries, the number of deaths due to traffic accidents is expected to increase by 80% between 2000 and 2020.

Data mining is the extraction of hidden predictive information from large databases and it is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It is a useful tool to address the need for shifting useful information such as hidden patterns from databases.

Data mining is supported by three technologies as follows

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Road accidents have been the major cause of injuries and fatalities in worldwide for the last few decades. A traffic collision occurs when a road vehicle collides with another vehicle, pedestrian, animal, or geographical or architectural obstacle. It can result in injury, property damage, and death.

Traffic control system is the area, where critical data about the society is recorded and kept. Using this data, we can identify the risk factors for vehicle accidents, injuries and fatalities and to make preventive measures to save the life. The severity of injuries causes an impact on the society.

Understanding the patterns of hidden data is very hard due to data accumulation. Organization keeps data on their domain area for maximum usage. Apart from the gathering data it is important to get some knowledge out of it. For effective learning, data from different sources are gathered and organized in a consistent and useful manner.

The periodic investigation of accident characteristics is vital to define target groups for further actions and measures and also helps decision makers to set rational policies and strategies to curb this problem. It includes characteristics of severity such as weather conditions, light conditions, road type, number of vehicles, casualty age, location of injury, degree of injury, pedestrian action, driver Age, driver sex, severity of accident and vehicle class.

The main objective of the research is to investigate the role of road-related factors in accident severity using RTA data and predictive models. Our two specific objectives include:

1) Exploring the underlying variables that impact accident severity.
2) Predicting accident severity using Random Forest algorithm.

II. RELATED WORK

Various studies have addressed the different aspects of RTAs, with most focusing on predicting or establishing the critical factors influencing injury severity. Numerous data mining-related studies have been undertaken to analyze RTA data locally and globally, with results frequently varying depending on the socio-economic conditions and infrastructure of a given location.

Hashem R.AL-Masaeid used two types of descriptive models to analyze the reason for RTAs- Relationship between no. of accidents and motorization &To estimate fatalities as a function of motorization level. These two models successfully explain variations in traffic accidents and fatalities. His study included the analysis of traffic accidents data from 1998 through 2007. Two safety policies were implemented to curb this problem: Application Temporary law (No.52,2007) which imposes penalty on driver's violation and Intensification of traffic police enforcement. Finally, enforcement of these laws and improvement of rescue medical services are considered as an essential part of safety policies.

To analyze the relationship between RTA severity and driving environment factors, **Sohn** and **Lee**

(2002) used various algorithms to improve the accuracy of individual classifiers for two RTA severity categories. 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 indicated that a clustering-based classification algorithm works best for road traffic accident classification in Korea.

To predict accident severity, **TibebeBeshah** and **Shandraw Hill** used certain classification algorithms. Classification models were built using decision tree, Naive Bayes and K-nearest neighbor classifier. They collected and cleaned traffic accident data, and tested a number of predictive models.Here classification is based on road related factors such as road orientation, road separation, road surface type and road surf condition. In this they observed that casualty frequencies were higher in accidents that occurred nearer to residential zones possibly due to higher exposure. Casualty rates among residents from relatively deprived areas were higher than those of affluent areas.

Tibebe Beshah, Dejene Ejigu, Ajith Abraham, Vaclav Snasel and Pavel Kromer used two approaches to identify relevant patterns and illustrate the performance of the techniques for the road safety domain. They are Classification and Adaptive Regression Trees (CART) and TREENET approach. While comparing two algorithms they all perform well in determining non-injury risk of an accident compared to injured risks. The TREENET predictive modeling technique performs better by exhibiting lower error rate, higher ROC score and greater prediction accuracy than CART. The results showed that the models could classify accidents with promising accuracy. These results can help in improving the road safety.

Khair S Jadaan,Muaath Al-Fayyad, and Hala F. Gammohused Artificial Neural Network (ANN) which is a novel approach and proved to be successful in solving engineering problems. Developing the ANN model for accident prediction involves in data collection which included the following input data such as: number of registered vehicles, population, total length of paved roads, gross domestic product. The model was validated and found to produce good results under Jordanian traffic conditions thus can be used with confidence to predict future traffic accidents on the national road network.

S. Krishnaveni and Dr.M.Hemalatha evaluated application of Naive Bayes, AdaBoostM1, PART, J48 and compares these algorithms' performance based on injury severity. The dataset which they considered has only drivers' records and does not include passengers' information. The attributes related to driver are vehicle class of driver or passenger casualty, driver age, driver sex, year of manufacture, severity of accident and vehicle class. The injury severity has three classes: Based on Accident, Based on Vehicle and Based on Casualty.

III. METHODS AND MATERIAL

The main objective of the proposed methodology is to build the prediction classification rules of the best performing model (J48, ANN, or Random Forest). This section explains the proposed research methodology to compare the performance of the J48, ANN, and Random Forest Classifier models and to use the most accurate predictive model.

Data Pre-Processing:

Data preprocessing is an important stage for handling the data before using it in the data mining

algorithms. This process involves various steps, including cleaning, normalization, feature selection, transformation.

Dataset Description:

Table-1: Dataset Description

Feature Name	Feature Description and Values
Number of Vehicles	Number of Vehicles involved in an accident
Number of Casualties	Number of Casualties involved in an accident
Road Type	1: Roundabout 2: One-way street 3: Dual carriageway 4: Single carriageway 5: Slip Road
Speed limit	The Speed limitation of the road where the accident happened
Light Conditions	1: Daylight 2: Darkness - lights lit 3: Darkness - lights unlit 4: Darkness - no lighting 5: Darkness - lighting unknown
Road Surface Conditions	1: Dry 2: Wet or damp 3: Snow 4: Frost or ice 5: Flood over 3cm. deep 6: Oil or diesel 7: Mud 8: Data missing or out of range
Weather Conditions	1: Fine no high winds 2: Raining no high winds 3: Snowing no high winds 4: Fine and high winds 5: Raining and high winds 6: Snowing and high winds 7: Fog or mist 8: Other 9: Unknown
Urban or Rural Area	1: Urban 2: Rural

Accidents are classified into 3 categories:
1.Fatal
2.Serious
3.Slight

Feature Selection:

Feature selection, also known as attribute selection or variable selection, is a process of selecting a subset of relevant features for using in model construction. We used Information Gain and Gain Ratio measures to rank the attributes and determine the most useful attribute, and accordingly, we determined different thresholds for the number of the most influential attributes to be used in the experiments. Then the used algorithms were applied to the dataset with these selected features, and the accuracies of them were compared and repeated this process with the multi thresholds to obtain the highest accuracies.

Classification Using Data Mining Algorithms:

After preprocessing step, Data Mining algorithms are performed on the dataset to find the best one in the prediction of the traffic accident severity by comparing the accuracies between them. Data mining has various tasks such as classification and prediction,

clustering, association rule mining. Classification techniques classify data into the predefined class label. Data classification is a two-step process. The first step is the learning phase where training data are analyzed to build a model (classification rules) that describes a predefined set of classes. The second phase is the classification phase where the accuracy of the model is estimated using test data.

J48 Decision Tree Classifier:

J48 is a simple C4.5 decision tree, it creates a binary tree.C4.5 builds decision trees from a set of training data which is like an ID3, using the concept of information entropy.

Artificial Neural Networks:

Artificial Neural Network is a sub-domain of artificial intelligence system. A neural network is a data-modeling tool and an information processingparadigm that represent complex relationships in a manner similar to the human brain. ANNs are known to be universal function approximations.

Random Forest Classifier:

A random forest consisting of a collection of tree structured classifiers (h(x, _k), k = 1. . .) where the _k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

Algorithm:

- Choose T number of trees to grow.
- Choose m number of variables used to split each node. m<<M, where M is the number of input variables, m is hold constant while growing the forest.
- Grow T trees. When growing each tree do.
- Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample.
- When growing a tree at each node select m variables at random and use them to find the best split.
- Grow the tree to a maximal extent and there is no pruning.

IV. EXPERIMENTATION AND RESULT EVALUATION

This section presents and discusses the experiments and the results for the three different classifiers (Random Forest, J48 and ANN (back-propagation)). Different comparisons and analysis were discussed in this section to see which of the three approaches provide better performance on prediction traffic accident severity. Accuracy, precision, recall and F1-measure measures were used in comparisons.

A. Dataset Splitting:

In the above section we explained the preprocessing steps for our dataset, after finishing preprocessing, we need to split this dataset into a Training dataset and Testing dataset, different ways are used to do this.

B. Experiments and Results:

The following results are for J48 decision tree, Artificial Neural Networks (ANN) and Random forest classifiers

1. J48 Decision Tree Classifier

The highest results of the J48 classifier using Down-Sample, Up-Sample, and Hybrid datasets were summarized in the table:

Table 2 - The overall highest results of performance measures for J48.

sampling	Evaluation Method	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Under sampling	Hold out	6	49.01	49.1	49.9	49.3
Over-sampling	10-fold CV	8	62.19	61.1	62.0	61.6
Hybrid	10-fold CV	8	66.24	63.5	69.0	69.0

2. ANN

The highest results of the ANN classifier using Down-Sample, Up-Sample, and Hybrid datasets were summarized in following table:

Sampling	Evaluation Method	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Under sampling	10-fold CV	6	48.01	48.1	48.4	47.8
Over-sampling	Hold out	8	50.19	49.1	51.1	50.0
Hybrid	10-fold CV	8	61.24	59.5	61.0	69.0

3. Random forest Classifier

All these results were achieved on the Hybrid sampling dataset and with Holdout method and 10-folds Cross Validation for each ANN, Random forest and J48 classifiers.

The highest results of the Random forest classifier using Down-Sample, Up-Sample, and Hybrid datasets were summarized in following table:

Table 4. The overall highest results of performance measures for Random forest classifier.

Sampling	Evaluation Method	No. of features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Under sa sampling	10-fold CV	6	49.01	49.1	49.2	49.0
Over-Sampling	Hold out	8	3.09	62.5	63.0	62.7
Hybrid	10-fold CV	8	1.24	9.5	71.0	79.0

The best result of ANN was on the Hybrid dataset, and the results were 61.4%, 50.193%, and 48.01% on Hybrid, Oversampling, and Under-Sampling datasets respectively.

The best result of J48 was on the Hybrid dataset, and the results were 54.8%, 47.4%, and 46.6% on Hybrid, Under-Sampling, and Over-Sampling datasets respectively.

The best result of Random forest classifier was on the Hybrid dataset, and the results were 78.9%, 62.5% and 49.8% on Hybrid, Oversampling and Undersampling

data sets respectively.

From the above analysis, it is clear that Random forest has the highest accuracy .In addition, the results of Precision, Recall, and F-Measure were poor for minor classes (Fatal and Serious) and bias to the major class (Slight) this due to the skewed distribution of data between classes on an imbalanced original dataset.

Random forest has many advantages over different classifiers. This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees. The random forest algorithm works well when there are both categorical and numerical features. It works well when data has missing values or it has not been scaled well (although we have performed feature scaling in this article just for the purpose of demonstration).

It is unexcelled in accuracy among current algorithms and it runs well on large data bases. It can handle thousands of input variables without variable deletion and also the learning is so fast. It computes proximities between pairs of cases that can be used in clustering, locating outliers or give interesting views of the data.

V. CONCLUSION

Road traffic accident constitutes a serious problem and prediction of its magnitude using reliable approaches has become a necessity. An accident prediction model was developed using the Random forest algorithm through analyzing the relationship between accidents and parameters affecting them for which data were available. In this paper, we collected and cleaned traffic accident data, attempted to construct novel attributes, and tested a number of predictive models. The outputs of the models were presented for analysis to domain experts for feedback.

Random Forest outperforms than other classification Algorithms instead of selecting all the attributes for classification. In this work, we extended the research to different cases such as Accident, Casualty and Vehicle for finding the cause of accident and the severity of accident.

VI. REFERENCES

[1]. Abdel-Aty, M., and Abdelwahab, H. 2003,"Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility". Accident Analysis and Prevention.

[2]. Akomolafe, T., and Olutayo, A. 2012. "Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways." American Journal of Database Theory and Application 1 (3): 26-38.

[3]. Al-Masaeid, H.R., (2009). Traffic accidents in Jordan. Jordan Journal of Civil Engineering, 3(4), pp.331-343.

[4]. Al-Zubi, A. S. A., (2010). Analysis of Vehicles Accidents in Amman City Using Spatial Data Mining and Visualization (Doctoral dissertation, The University Of Jordan).

[5]. Beshah, T. and Hill, S., (2010), Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. In AAAI Spring Symposium: Artificial Intelligence for Development.

[6]. Beshah, T., A. Abraham, et al. (2005). "Rule Mining and Classification of Road Traffic Accidents Using Adaptive Regression Trees." Journal of Simulation 6(10-11).

[7]. Beshah, T., Ejigu, D., Abraham, A., Snasel, V. and Kromer, P., (2013). Mining Pattern from Road Accident Data: Role of Road Users Behaviour and Implications for improving road

safety. International Journal of Tomography and Simulation, 22(1), pp.73-86.

[8]. Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

[9]. De Ona, ˜ J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention 43, 402–411.

[10]. Effati, M., Rajabi, M.A., Hakimpour, F. and Shabani, S., (2014). Prediction of crash severity on two-lane, two-way roads based on fuzzy classification and regression tree using geospatial analysis. Journal of Computing in Civil Engineering, 29(6), p.04014099.

[11]. El Tayeb, A. A., Pareek, V., Araar, A., (2015). Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai. International Journal of Soft Computing and Engineering (IJSCE), 5(4).

[12]. Helai, H., Chor, C.H., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54.

[13]. Jadaan, K.S., Al-Fayyad, M. and Gammoh, H.F., (2014). Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN). Journal of Traffic and Logistics Engineering, 2(2).

[14]. Krishnaveni, S., and Hemalatha M., (2011). A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7), pp.40-48.

[15]. Kunt, M.M., Aghayan, I. and Noii, N., (2011). Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. Transport, 26(4), pp.353-366.

[16]. Mohamed, E.A., (2014). Predicting Causes of Traffic Road Accidents Using Multi-class Support Vector Machines. In Proceedings of the International Conference on Data Mining (DMIN) (p. 1).

[17]. Obaidat, M.T., and Ramadan, T.M., (2012). Traffic accidents at hazardous locations of urban roads. Jordan Journal of Civil Engineering, 6(4), pp.436-447.

[18]. Olutayo, V.A., and Eludire, A.A., (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. International Journal of Information Technology and Computer Science (IJITCS), 6(2), p.22.

[19]. Ossenbruggen, P.J., Pendharkar, J. and Ivan, J.2001, "Roadway safety in rural and small urbanized areas". Accidents Analysis and Prevention, 33 (4), pp. 485– 498.

[20]. Perone, C.S., (2015). Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil. arXiv preprint arXiv:1502.00245.

[21]. Prediction Based on Neural Network." Presented at the 2nd International Conference on DigitHuilin, F., and Yucai, Z. 2011. "The Traffic Accidental Manufacturing & Automation.

[22]. Sohn, S. and S. Hyungwon (2001). "Pattern recognition for a road traffic accident severity in Korea." Ergonomics 44(1): 101-117.

[23]. Srisuriyachai, S. (2007). Analysis of road traffic accidents in NakhonPathom province of Bangkok using data mining. Graduate Studies. Bangkok, Mahidol University.

[24]. WondwossenMulugeta. 1999, "Correlates of car traffic accident: the case of Addis Ababa in 1990". Addis Ababa University, Addis Ababa.

[25]. Yousif, J.H., and AlRababaa, M.S., (2013). Neural Technique for Predicting Traffic Accidents in Jordan. Journal of American Science, 9(11).

Cite this article as :

S. Ramya, SK. Reshma, V. Dhatri Manogna, Y. Satya Saroja, Dr. G. Sanjay Gandhi, "Accident Severity Prediction Using Data Mining Methods", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 528-525, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195293
Journal URL : http://ijsrcseit.com/CSEIT195293