Improved Accuracy of Calculation of Vehicle Crash Severity in Highways using Random Forest over Decision Tree Algorithm

Vignesh.S 1, Sashi rekha K 2

1Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.
2Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

Aim: To improve the accuracy rate of vehicle crash severity in highways using Random forest over Decision tree. Materials and Methods: Random forest and Decision tree with sample size of (N=10) is executed with varying training and testing splits for calculating the accuracy for accident crash severity with g power as 75%, threshold 0.000 and confidence interval 95%. The performance of the classifiers are evaluated based on their accuracy rate using accident severity dataset. Results: The accuracy for calculating accident crash severity in Random Forest(91%) and Decision Tree (85%) is obtained(P<0.005).Conclusion: Prediction of accident crash severity using Random Forest (RF) algorithm appears to be significantly better than Decision Tree (DT) with improved accuracy.

Keywords: crash severity, Novel Random Forest Algorithm, Decision Tree Algorithm, Machine Learning, Artificial Intelligence.


INTRODUCTION

Accidents are naturally unspecified happening incidents. The increase of vehicles in these years has also increased the number of accidents and deaths. To predict the severity of accidents and to prevent death and reduce injury level this method is implemented (Ahmadi et al. 2020). The importance of crash severity prediction in recent years has helped hospitals provide proper medical care as fast as possible when an accident occurs and also to amend road safety (Zhang et al. 2018). This research helps to predict the crash severity for transportation safety planners so that hospitals and agencies can provide emergency care with artificial intelligence (Iranitalab and Khattak 2017). It is useful in reducing the repercussion of traffic crashes (Assi et al. 2020).

A lot of research has been performed on crash severity prediction using machine learning algorithm. 1761 research articles were published in Google Scholar and 37 articles were found in IEEE Xplore. In this work RF, Adaboost, GBDT, Logistic Regression is used to help the traffic management department to predict the accuracy of crashes with regarding to road infrastructure (Tang et al. 2019). This proposed system can be extend the application for studies on the condition of environments in two way rural highways because of the concerning impacts of crash severity using multiple Decision Trees (Abellán, López, and de Oña 2013). In this various algorithms like ANN with 90% accuracy, SVM with 89% and RF with 92% has been used. During travel we can check constantly for crashes occurring via the probe vehicle based on location and speed it can direct to take secondary route to avoid furthermore accidents with assist of artificial intelligence (Dogru and Subasi 2018). SVM model is used in this paper to evaluate the sensitive impact of explanatory variables to measure the severity of the crash (Li et al. 2012). The Random Forest Algorithm proved most efficient based on the accuracy and is used to predict the correct accident severity of the dataset (Geyik and Kara 2020).
Our institution is passionate about high quality evidence based research and has excelled in various fields (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyaamoorthy et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilaraas, and Rajeshkumar 2020; Kamath et al. 2020). Based on the literature survey it can be concluded that the existing accident severity prediction model was not able to accurately determine severity of crash in terms of accuracy. In the proposed work the lack of accuracy in predicting the accident crash severity is analyzed and improved using machine learning algorithms such as Novel Random Forest algorithm and Decision Tree Algorithm.

MATERIALS AND METHODS

The study setting of the proposed work is done in the Image Processing Lab, Department of Computer science and Engineering at Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (SIMATS). The number of groups used for the study is two group classification algorithms. The Group 1 is Novel Random Forest Algorithm and Group 2 is Decision Tree Algorithm. Using clinical analysis (Kane, Phar, and BCPS n.d.), the analysis of sample size of N = 10 has been carried out with confidence of 95%. The input dataset is collected from kaggle.com (Accident Severity). It consists of two dataset namely test and train. Both the datasets have 17 attributes. Table 1 Contains the description of the dataset, the attribute which are present in the dataset incorporates like collision ref no, policing area, day, month, week, hour, weather conditions etc. Totally it consists of 1550 rows in the datasets which has null and duplicate values.

Random Forest Algorithm

Novel Random Forest Algorithm extensively used for both regression and classification problems. This algorithm can be used in various places such as banking, prediction works, health, stock markets, artificial intelligence etc. To get more accurate and stable estimates the random forest will create forest like structures and combines them. Subsets from both datasets and attributes are selected arbitrarily and gets trained using this method overfitting of data can be lowered. This algorithm takes lower training time than many other algorithms on large datasets with maintaining precision of the accuracy when a huge part of data is not present. Table 2 contains the pseudocode of Random Forest Algorithm.

Decision Tree Algorithm

This algorithm is used in predictive modeling which can be used for statistics, data mining, artificial intelligence etc. It will be like branches where the target variables are done for observations of a pattern. If the outcome has target variable or any classification values are known as classification trees. If the values are uninterrupted then they are called regression trees. Then further this tree is dividing the subsets as node and then to establish successor child which will occur recursively. Thus the prediction is made.

\[
\text{Entropy} = \sum_{x \in X} - p(x)E(a) \\
\text{Information Gain} = \text{Gain}(A, S) = E(S) - \sum_{a \in T} p(a)E(a)
\]

Here in (1) and (2) S is the dataset, class set is given by X, ratio of elements in x class is represented by p, the subset of data created from database S is represented by T. Table 3 contains the pseudocode of Decision Tree Algorithm.

Statistical Analysis

The algorithms are run on a Personal Computer with 64-bit Operating System, 8GB RAM with steady internet connection and software such as Collab Notebook and python 3.9 for executing these algorithms. The independent variable are junction detail, junction control whereas the dependent variables are policing area, week day of collision, hour of collision etc. The software used here for statistics is SPSS version 26. The data set is prepared by using 5 iterations each of both Random Forest and Decision Tree. The testing variables are accuracy and loss whereas the group ID is given as grouping.

RESULTS

The mean accuracy and loss values using T-test for both the algorithms along with the standard deviation is shown in Table 6. Here RF and DT classifiers are used. The performances of the classifiers are measured by accuracy value. The Dataset is split into training and testing data to find accuracy. Where Table 4 contains the accuracy for Random Forest classifier with N = 5 and Table 5 contains the accuracy for Decision Tree classifier with N = 5.
Random Forest Classifier input is taken from the datasets as a class form and gives the output accuracy of 91%. Decision tree classifier input is taken as class form from the data set and gives the output accuracy of 85%. Table 8 shows the accuracy values for both Random forest and Decision tree algorithm. It can be seen that Random forest has given better results of accuracy than Decision Tree. From Table 7 it can be observed that Random Forest has better significance value than Decision Tree with a value of $p = 0.000$. Accuracy and the Loss for both the algorithms are presented in a bar graph in Fig. 1.

**DISCUSSION**

Random forest as convincingly appears better than Decision Tree with improved accuracy as referred in Table 6. The Novel Random Forest classifier shows some difference in terms of accuracy score speed and performance in comparison with Decision Tree.

Comparison between each algorithm’s performance in terms of accuracy, and loss proved that Random Forest is the best overall. The outcome of this paper is similar to (Nour et al. 2020) where they found that Random Forest Algorithm shows better performance with 73% accuracy than Decision Tree. Hence Random Forest can be used to predict accident crash severity efficiently. (Elyassami, Hamid, and Habuza 2020) In this work the experiment was carried out using regressive testing of all possible variations that are there for the given models. The Random Forest method involved the use of various tree depth levels along with the use of confidence vote and Gini index criteria. It resulted in an accuracy of 72.98. In the proposed System it can be found that the staging framework has an enormous number of hyper parameters that are to be calibrated. The classification models can be enhanced inorder to improve the features of classification. (Wahab and Jiang 2019) In this work the Random Forest model showed the best prediction performance and surpassed the traditional MNLM methods. (Chen and Chen 2020) paper showed that the results produced by Random forest in terms of accuracy and specificity are better compared to that of other machine learning algorithms. The study was done using 15 original values and 10 significant variables. Similarly in (Ji and Levinson 2020) this work RF shows best performance for predicting the severity using its iterative improving algorithms.

Most of the previous work done is based on the data from the specific dataset. In the proposed work the lack of accuracy in predicting the accident severity is improved by machine learning algorithms. In future more attributes in the dataset can be included and images can also be used as datasets for better accuracy. Artificial intelligence can also be used in future to predict the severity of the crash.

**CONCLUSION**

The Calculation of accuracy of accident crash severity in highways is carried out using the dataset obtained from Kaggle. It is done by Random Forest and Decision Tree. The accuracy of Random Forest classifier is 92% and Decision Tree with 85% accuracy. The accuracy of predicting the accident crash severity is more in Novel Random Forest than Decision Tree.

**DECLARATIONS**

**Conflict of interests**
No conflict of interest in this manuscript.

**Authors Contributions**
Author SV was involved in data collection, data analysis, manuscript writing and Author SRK was involved in conceptualization, data validation, and critical review of manuscript.

**Acknowledgment**
The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

**Funding:** We thank the following organizations for providing financial support that enabled us to complete the study.

1. Soft Square Solutions, Palavakkam, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES


Sashi rekha K, et.al: Improved Accuracy of Calculation Of Vehicle Crash Severity in Highways using Random Forest over Decision Tree Algorithm

Tables and Figures

**Table 1. Dataset Description**

<table>
<thead>
<tr>
<th>Column</th>
<th>Values (For categorical variables)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collision_Ref_No</td>
<td>Multiple Values Present</td>
<td>String, Categorical</td>
</tr>
<tr>
<td>Weekday_of_Collision</td>
<td>Multiple days present</td>
<td>String, Categorical</td>
</tr>
<tr>
<td>Special_condition_at_site</td>
<td>1 (Yes), 0 (No)</td>
<td>Numerical, Categorical</td>
</tr>
<tr>
<td>Ped_Crossing_HC</td>
<td>1 (Yes), 0 (No)</td>
<td>Numeric, Categorical</td>
</tr>
<tr>
<td>Policing_Area</td>
<td>Multiple Values present</td>
<td>String, Categorical</td>
</tr>
<tr>
<td>Hour_of_Collision</td>
<td>Multiple Values</td>
<td>Numeric, Categorical</td>
</tr>
</tbody>
</table>

**Table 2. Pseudocode for Random Forest Algorithm**

Input
- Accident-severity (train and Test) Dataset

1. Initialization // input_attributes
2. The Random Forest Classifier is fit on the training set
3. Predict the records based on scaled values.
4. The features are then trained, tested and summarized
5. Predict the precision score

Output: Prediction of Accuracy

**Table 3. Pseudocode for Decision Tree Algorithm**

Input
- Accident-severity (train and Test) Dataset

1. Initialization // input_attributes
2. Import Decision classifier
3. The Decision tree Classifier is fit on the training set
4. Predict the records based on scaled values
5. Precision values scores are predicted

Output: Prediction of Accuracy

<table>
<thead>
<tr>
<th>Table 4. Random Forest Accuracy and Loss for N = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterations</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5. Decision Tree Accuracy and Loss for N = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterations</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

<p>| Table 6. T-Test Group Statistics with Mean, Std.Deviation, Std.Error Mean and Confidence = 95% |</p>
<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>5</td>
<td>91.8680</td>
<td>.28084</td>
<td>.12559</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>5</td>
<td>85.2860</td>
<td>.48531</td>
<td>.21704</td>
</tr>
<tr>
<td>Loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>5</td>
<td>8.1320</td>
<td>.28084</td>
<td>.12559</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>5</td>
<td>14.5240</td>
<td>.13297</td>
<td>.05946</td>
</tr>
</tbody>
</table>

| Table 7. Independent Sample T-Test is applied for the data set fixing confidence interval as 95% and level of significance as 0.05 |

Sashi rekh K, et.al: Improved Accuracy of Calculation Of Vehicle Crash Severity in Highways using Random Forest over Decision Tree Algorithm

<table>
<thead>
<tr>
<th>CLASSIFIER</th>
<th>ACCURACY(%)</th>
<th>LOSS(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>91.70%</td>
<td>8.30%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>85.49%</td>
<td>14.51%</td>
</tr>
</tbody>
</table>

Table 8. Comparison of the Random Tree Algorithm and Decision Tree Algorithm with their accuracy
Fig. 1. Comparison of Random Forest algorithm and Decision Tree in terms of mean accuracy. The mean accuracy of Random Forest is better than Decision Tree. X axis (Groups): RF VS KNN algorithm, Y axis: Mean accuracy of prediction +/- 2SD.