International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1953138

Analysing Road Accident Criticality using Data mining

Shahsitha Siddique V*, Nithin Ramakrishnan
Department of Computer Science and Engineering, MEA Engineering College, Perinthalmanna, Kerala, India

ABSTRACT

Road transport is one of the most vital forms of transportation system, connecting both long and short distances in our country. There are several attributes, which affect the intensity of a road accident like speed of the vehicle, road conditions, time of the accident etc. Analysing these attributes gives an idea about the factors lead to the severity of the accident. Data mining is a method to analyse huge amount of traffic data in an efficient manner, which gives the factors, affect the road accidents. Several machine learning algorithms can be used to find the relation between traffic attributes the lead to the severity of the accidents. In this work, we use three methods for predicting accident criticality. First, Naive Bayesian Classifier is used to get the accident severity based on Bayes rule. Then, Decision Tree classifier is used for same purpose for accident severity calculation. Finally K-Nearest Neighbour(KNN) classifier is employed for severity calculation. The accuracy of the algorithms are compared and it is found that KNN performs better than the other two algorithms employed. The major aim of the work is to find the accident severity. Also the work aims to reduce road accidents by giving awareness to public using the above method.

Keywords: Data mining, Naive Bayes, Decision Tree, K-Nearest Neighbor, Classification.

I. INTRODUCTION

Data mining (sometimes referred to as data or knowledge discovery) is the way to examine and condense information from alternative perspectives into useful data. Data mining is the computational process of discovering designs in enormous information collections, including methods of man-made consciousness convergence, AI (Artificial Intelligence), perspectives, and database frameworks. Accidents are unpredictable and happen in different circumstances. Therefore, it can be helpful to understand the factors that contribute to an accident in stopping it.

The general objective of the data mining method is to obtain information from a set of data and convert it for further use into an understandable framework [1]. The real job of information mining is to semi-automatically or automatically analyse big amounts of information to obtain earlier unknown, interesting patterns such as data record organizations, uncommon documents and dependencies principles and should provide authentic contribution to knowledge.

Classification of accidents is a standardized technique of grouping the causes of an accident into categories, including the root causes. Classification of accidents is primarily used in aviation, but can be extended to other fields such as rail or health care. The most prevalent cause of road accidents is distracted driving, leading in more crashes each year than speed, drunk driving, and other significant causes of accidents [5]. Analysis of accidents is very crucial because the link between the different types of features that add to an

incident can be exposed. Compared to other data on the real globe, highway, traffic and aircraft accidents are of a distinct nature, as accidents are uncertain. Analysing different accident information collection can provide information on the contribution of these features that can be used to deteriorate the frequency of the accident.

The research's overall goal is to explore the role of road-related variables in the severity of accidents using classification models. The primary commitments of this paper are as follows:

• Exploring the underlying (particularly road-related) factors affecting the seriousness of car accidents.

• Predicting accident seriousness using separate data mining techniques.

• Comparison of standard models for this assignment.

The scope of this work is intended to analyze the database and take preventive measures to avoid fatal accidents using data mining. India has the world's second biggest highway network. Road accidents occur quite commonly and every year they claim too many life. To avoid them, the root cause of road accidents must be discovered [8]. In order to recognize possible hidden interactions and links between different variables influencing road accidents with deadly effects, the appropriate data mining method must be implemented to collected information sets representing occurring road accidents. The objective of this job is to reduce the number of road accidents caused by many variables or situations. In this extract interesting patterns and analyze using data mining algorithms. This paper discusses some of the classification models to predict the severity of traffic accidents injury.

II. METHODS AND MATERIAL

The methodology allotted for this work involves a sequence of steps, data extraction it is the first and the foremost step where large amount of data like accident severity, road crash reports, road traffic accident reports as produced under the transport ministry are pooled and collected. Then these data collected from various sources are then stored in a database. Now these collected and stored data are subjected to pre-processing where data cleaned, transformed to reduce noisy data and also to fill missing values. Then these data is stored in database and then further searching and analyzing can be done using it. Finally data visualization is done which helps in communicating data or information by displaying it as visual objects.

A. Data Set

The datasets for training the classifier is obtained from Government of Telangana Transport Department. The dataset consist of around 9000 samples of 16 traffic attributes like Time of accident, area, speed, season, lighting condition etc. [1]. Here are 3 classes under the Collision Severity {1,2,3}. The evaluation is the accuracy metric. The data collected is not in the format required to analyze it in an unstructured format, so the data must be preprocessed. It is important to eliminate missing values from the dataset since there were no values that could negatively impact the real performance. Hence, missing values must be destroyed in the dataset. After removing missing values from the dataset, then cleaning duties. The datasets can be downloaded from (https://www.kaggle.com/c/accident-severity).

The "Fig. 1" shows the system architecture where, the process begins load the data set for the entire data then these data are subjected to pre processing where data cleaning is held and unwanted noise is filtered off. Then after these procedures the training data is obtained and also we add the test data to the classification model to get the prescribed output.

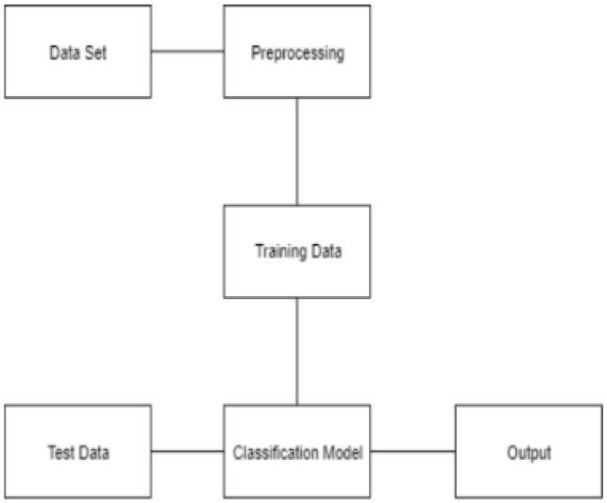

Figure 1. System Architecture for Accident Criticality Prediction

B. Data Pre-processing

Data preprocessing is one of the most significant functions in information mining. Data preprocessing primarily involves removing noise, handling missing values, removing irrelevant characteristics to make the information ready for assessment. In this step, our goal is to pre-process the accident information in order to make it suitable for assessment.

C. Classification Modelling

This section describe the methods selected classification methods for accident severity prediction. First, using Naive Bayes classifier classification is carried out. Then, the output of is displayed, for a given set of attributes which predicts that whether the accident would be critical or noncritical. Then, to obtain association the same data is subject to classification using Decision Tree. Accidents can be analysed Using this data. Naive Bayesian classifier based on Bayes rule is used to get the severity. Decision tree classifier is another classifier, which gives good result for accident severity calculation. Finally K-Nearest Neighbor (KNN) classifier in employed for severity calculation. The accuracy of the algorithms are compared and it is found that KNN performs better than the other two algorithms employed.

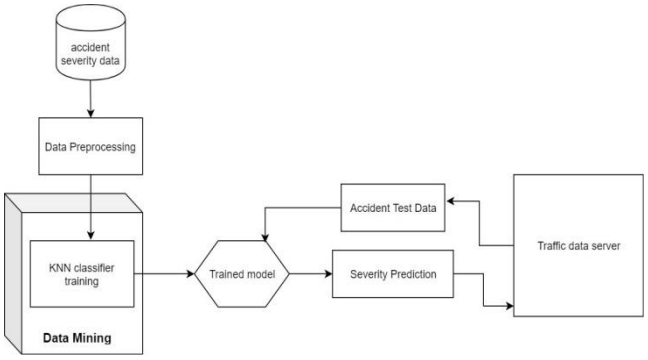

Figure 2. Block Diagram of the Methodology

The principle work process of our framework is as appeared in "Fig. 2", which include few stages, like Data pre-processing, Classification modeling and accident criticality prediction. The dataset obtained is pre-processed first. Some of the attributes missing in some cases so pre-processing is necessary to avoid confusion in training the dataset. The pre-processed data is used for training a K Nearest Neighbour (KNN) Classifier. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour. The classifier is trained in less than a minute. The trained classifier is then do for testing the data. The test data is also obtained from the same website. The evaluation metric used are Accuracy, Precision and Recall. The results shows that KNN outperforms the other classifiers like Naive Bayesian and Decision Tree classifier etc.

Bayes Theorem

Bayes Theorem finds the likelihood of an occasion happening given the likelihood of another occasion that has just happened. The Naive Bayes Classifier method depends on the supposed Bayesian

hypothesis and is especially fit when the dimensionality of the information sources is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical.

Naive Bayes Classifier

Naive Bayes is a probabilistic classifier based on Bayes theorem. It assumes variables are independent of each other. The algorithm is easy to build and works well with huge data sets. It has been used because it makes use of small training data to estimate the parameters important for classification. Bayes Theorem states the following:

$$P(c|x) = \frac{P(X|C)P(C)}{P(X)}$$

$$P(c|x) = P(x_1|c)*P(x_2|c)*\ldots P(x_n|c)*P(c)$$

Where P(c|x) is the posterior probability of class (target) given predictor (attribute). P(c) is the prior probability of class. P(x|c) is the likelihood which is the probability of predictor given class. P(x) is the prior probability of predictor.

Algorithm 1 Naive Bayes Classifier Algorithm

1. Calculate the preceding probabilities for each attribute class.
2. Calculate the likelihood of proof going into the denominator.
3. Calculate the probability of evidence going into the numerator.
4. Use the Bayers rule to calculate the probability of a specific attribute.

Decision Tree Classifier

Decision tree classification could be an ordinarily used data processing technique for establishing classification systems supported multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non parametric and can with efficiency trot out massive, difficult datasets while not imposing a sophisticated constant structure. When the sample size is massive enough, study information may be divided into coaching and validation data sets. Classification efficiency is the proportion of right predictions to complete prediction.

Algorithm 2 Decision Tree Classifier Algorithm

1. Consider the entire training set as the root at the start.
2. It is preferred that the feature values are categorical. If the values are continuous, they will be discrete before the model is built.
3. Recursive distribution of documents of attribute values.
4. For ordering characteristics as root or internal node, we use statistical methods.
5. Start with all the root node-related training cases.
6. Use the gain details to choose the attribute with which to label each node.
7. There should be no root-to-leaf route twice with the same discrete attribute.
8. Build each sub-tree on the sub-set of training cases that would be categorized in the tree down that route.
9. If all or all of the favorable or negative training cases stay, indicate that node is yes or no.
10. If there are no attributes remaining, label the training cases remaining at that node with a majority vote.

11. If no cases stay, label the parent training cases with a majority vote.

K- Nearest Neighbor Classifier

The k-Nearest Neighbor calculation (KNN) is a non-parametric strategy utilized for characterization and relapse in example acknowledgment. In the two cases, the information comprises in the component space of the closest k preparing models. The purchase changes depending on whether KNN's arrangement or relapse is used. K-Nearest Neighbor is one of the most significant but valuable framework estimations in data mining. It is a piece of the administered learning space and is seriously connected in example acknowledgment, information mining and interruption discovery.

Algorithm 3 K-Nearest Neighbor classifier

Let m be the amount of samples of training data. Let p be a point unknown.
1. Store training samples in a information point array[]. This implies that each component of this array constitutes a tuple (x, y).
2. For i=0 to m ;
3. Calculate the range of Euclidean d(arr[i], p).
4. Make set S of the lowest distances acquired. Each of these distances corresponds to an already classified data point.
5. Return the S majority label.

The accuracy of the algorithms are compared and it is found that KNN performs higher than the opposite two algorithms used. In KNN classification, the output may be a category membership. Associate object is assessed by a plurality vote of its neighbors, with the article being assigned to the category commonest among its k nearest neighbors (k may be a positive number, generally small.). If k=1, then the article is solely appointed to the category of that single nearest neighbor. The classifier is trained in but a second. The trained classifier is then doing for testing the information. The take a look at information is additionally obtained from identical web site. The analysis metric used are Accuracy, Precision and Recall. The results shows that KNN outperforms the opposite classifiers like Naive Bayesian and Decision Tree classifier etc.

C. Model Implementation

The models are implemented using MATLAB 2018. MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and proprietary programming language developed by Math Works. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C, Java, Fortran and Python. Advantages of MATLAB are Ease of use, Platform independence, Predefined functions and Plotting.

III. RESULTS AND DISCUSSION

The Result section shows that confusion matrix for each classifier is obtained. Using these values accuracy, precision and Recall is calculated. The proposed algorithm KNN performs better than the other two algorithms employed. Accident criticality or collision prediction gives more accurate results using proposed algorithm. In KNN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. The classifier is trained in less than a minute. The trained classifier is then do for testing the data. The test data is also obtained from the same website. The evaluation metric used are Accuracy, Precision and Recall classiffier. The results shows that KNN outperforms

the other classifiers like Naive Bayesian and Decision Tree classifier etc. The algorithms are used to rank it according to the Accident Severity class, where there are three values, namely low, medium, severe as values allocated as {1,2,3}.

B. Analysis of Results

This section includes discussion and identification outcomes gathered from different studies. The investigation center around the proposed algorithm and analysis of result.

C. Analysis of Proposed Algorithm

Table 3.1. Shows the results obtained for comparison of different classification algorithms. The result shows that KNN performs better than Naive bayes and decision tree algorithms.

From the table accuracy of NB Classifier obtained is 88% and that of DT Classifier gives 95% and using our proposed algorithm gives the accuracy of 99.9%. Precision value for NB classifier is 91% and that of DT classifier gives 97% and by KNN we can achieve it by 99.9%. Recall value using NB classifier is 98% and by DT Recall is 99%, and by proposed method, it obtained as 99.9%.

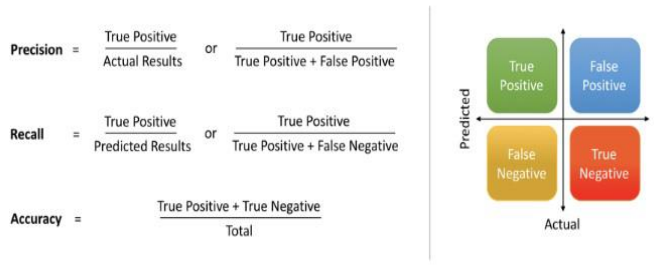

Figure 3. Accuracy Recall and precision calculation

As shown in "Fig. 3" accuracy is calculated with the help of confusion matrix. Confusion matrix is a table often used to define the performance of a classification model on a collection of test information. It contains 4 values namely, True Positive (TP) which is the correctly predicted event values, False Positive(FP) which is the incorrectly predicted event values, True Negative(TN) which is the correctly predicted no-event values, and False Negative(FN) which is the incorrectly predicted no event values. Accuracy of a model can be calculated by using these values obtained from the confusion matrix.

TABLE I
COMPARISON RESULTS FOR DIFFERENT CLASSIFICATION TECHNIQUES

Evaluation	NB(%)	DT(%)	KNN(%)
Accuracy	88%	95%	99.9%
Precision	91%	97%	99.9%
Recall	91%	99%	99.9%

D. Performance of Proposed Classifiers

In this section experimental comparison of Naive Bayes and Decision Tree and KNN done based on the performance vectors. It is statistical performance evaluation of classification tasks and contains list of performance criteria values. The performance vector-containing list of performance criteria values. Accuracy refers to number of correct predictions or how precise the dataset is being classified. The accuracy Precision and recall of KNN obtained is 99.9%.

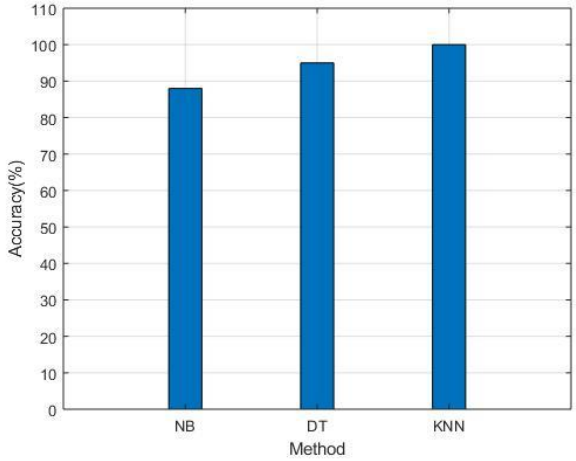

Figure 4. Accuracy comparison graph for different classification methods

"Fig. 4" Shows that the results that is the accuracy for NB classifier is 88% and that of DT got 95%. KNN classifier gives better accurate result than other classification methods that is 99.9% road accident criticality or collision predicted.

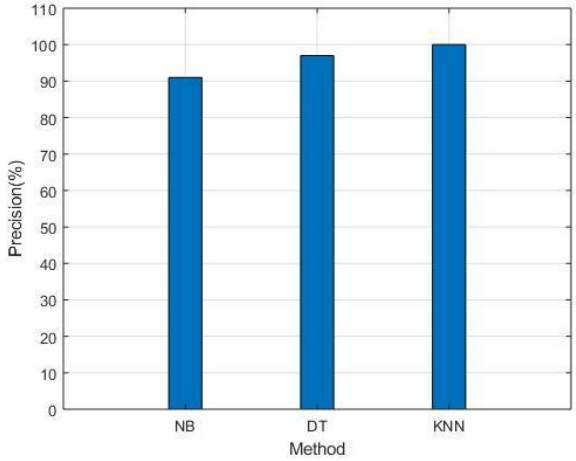

Figure 5. Precision comparison graph result for different classification methods

From "Fig. 5" the results shows that prediction value for NB classifier is 91% and that of DT classifier got 97%. KNN classifier gives better accurate result than other classification methods that is 99.9% road accident criticality or collision predicted.

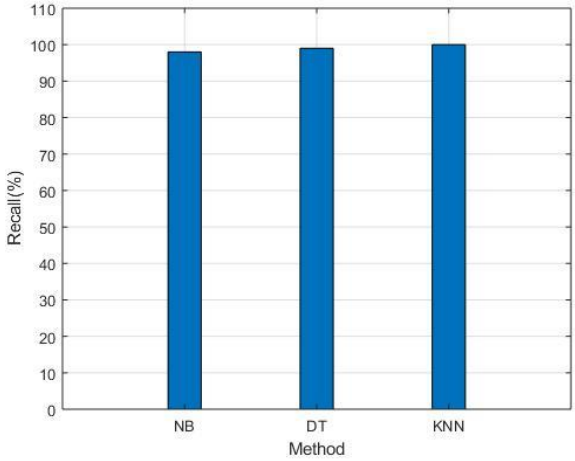

Figure 6. Recall comparison graph for different classification methods.

Here the result graph from "Fig. 6" shows that recall value for NB classifier is 98% and that of DT classifier got 99%. KNN classifier gives better accurate result than other classification methods that is 99.9% road accident criticality or collision predicted.

IV. CONCLUSION

In this work, the accuracy of K-Nearest neighbor (KNN) classifier is found to be more than that of Decision Tree algorithm, and Naive Bayes algorithm hence, KNN Algorithm can act as an optimal algorithm for this accident criticality prediction system which serves the purpose of predicting the criticality of an accident with the given attributes. It proves to be optimal when lesser attributes are taken into consideration. The major advantage of the work is it helps to give the peoples a warning about the possible accidents. Therefore, it works as an accident preventive system. The proposed work is compared with existing classifiers like Naive Bayesian, Decision Tree Classifier etc. and found to be the best. The parameters used to detect the performance of the classifiers are accuracy, precision and recall. KNN provides best value for all the parameters used. The conventional accident preventive system entirely depends on the traffic signboards. This is not enough

for preventing accidents. For example, the speed of the vehicle is limited in particular locations but the speed of the vehicle should be changed if the atmospheric condition is changed. In this work we propose an automated accident prevention warning system. By finding the accident severity using different parameters we can actually predict whether accidents likely to be happen or not. This is useful for the traffic departments for giving warning to people under different scenarios.

In future, this can be incorporated with road traffic signboards to prevent accidents in a more efficient way and in case of large amount of data set deep learning based Neural Network may use for accurate accident severity prediction results due to the reason of huge running time for our work.

REFERENCES

[1] Ramachandiran, V. M., PN Kailash Babu, and R. Manikandan. "Prediction of Road Accidents Severity using various algorithms." *International Journal of Pure and Applied Mathematics* 119.12 (2018): 16663-16669.

[2] Kwon, Oh Hoon, Wonjong Rhee, and Yoonjin Yoon. "Application of classification algorithms for analysis of road safety risk factor dependencies." *Accident Analysis & Prevention* 75 (2015): 1-15.

[3] Tavakoli Kashani, Ali, Afshin Shariat-Mohaymany, and Andishe Ranjbari. "A data mining approach to identify key factors of traffic injury severity." *PROMET-Traffic&Transportation* 23.1 (2011): 11-17.

[4] Montella, Alfonso, et al. "Data-mining techniques for exploratory analysis of pedestrian crashes." *Transportation research record* 2237.1 (2011): 107-116.

[5] Kumar, Sachin, and Durga Toshniwal. "A data mining framework to analyze road accident data." *Journal of Big Data*2.1 (2015): 26.

[6] Shanthi, S., and R. Geetha Ramani. "Feature relevance analysis and classification of road traffic accident data through data mining techniques." *Proceedings of the World Congress on Engineering and Computer Science*. Vol. 1. 2012.

[7] Chen, Zhuo, Xiaoyue Cathy Liu, and Guohui Zhang. "Non-recurrent congestion analysis using data-driven spatiotemporal approach for information construction." *Transportation Research Part C: Emerging Technologies* 71 (2016): 19-31.

[8] Li, Liling, Sharad Shrestha, and Gongzhu Hu. "Analysis of road traffic fatal accidents using data mining techniques." *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 2017.

Cite this article as :

Shahsitha Siddique V, Nithin Ramakrishnan, "Analysing Road Accident Criticality using Data mining", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 408-415, May-June 2019. Available at doi : https://doi.org/10.32628/CSEIT1953138
Journal URL : http://ijsrcseit.com/CSEIT1953138