Prediction of Road Accidents Severity using various algorithms

V M Ramachandiran, P N Kailash Babu, R Manikandan

School of Computing, SASTRA Deemed to be University, India.

ABSTRACT:

When a road accident occurs, there are several attributes that affect the intensity of the accident, like road conditions, weather conditions etc. The analysis of these accidents helps in identifying the features that lead to the criticality of the accident. In most cases, road accident analysis is done using Data Mining techniques, through which the factors that affect the severity of the accident can be analyzed. We have used classification algorithms to obtain the relation between various attributes that lead to the criticality of the accident. Firstly, Naïve Bayes classification algorithm is used to obtain the criticality of accident for the given attribute values. Then, Decision Tree algorithm is used to perform the same function. The accuracy of both the algorithms are compared so that the better algorithm can be used. The main aim of this project is to obtain a relationship between the factors leading to an accident and the criticality of the accident. The main contribution of this paper is to propose a tool to predict whether an accident which occurs for the given parameters is critical or non-critical.

Keywords: Data mining, Naïve Bayes, Decision Tree, Prediction, Classification.

1. INTRODUCTION:

Transportation plays a vital role in the economic success of a country. Majority of the places in the country is connected by roads. Even places which are not connected through rail routes are connected via roads. Several people use roadways to travel from one place to another. However, with the growing population, the number of people using roadways for transportation has increased. Roads carry more volume of traffic than it was actually designed to carry. This results in an increase in the number of vehicles on road. This has in turn increased congestion and traffic, vehicle crashes, etc. Accidents are undesirable damages that are caused unexpectedly. In India, Road and Traffic Accidents pose a serious issue. According to a report by MORTH [1], 0.4 Million accidents are reported every year. Accidents are unpredictable and they occur in various situation. Hence, understanding the factors that lead to an accident can prove to be useful in preventing it. Tremendous efforts have been taken to improve road safety. Traditional statistical methods cannot be used in such systems as the data generated is in large volumes, which demands the use of data mining algorithms for computation.

2. LITERATURE SURVEY:

MORTH mentioned that every year in India, 0.4 million accidents are reported. This makes India a country with a large accident rate [1]. It is said that data mining techniques like classification can be used for analysis of road accidents [2]. Classification and Regression Tree (CART) model has been used to analyse various factors that lead to an accident [3]. Naïve Bayes and Decision Tree classification algorithms have been used to analyse factors related to road safety. Hence, classification algorithms can be used to analyse factors leading to the accident [4]. Decision trees and association rule mining have been used to analyse pedestrian and two-wheeler accidents. Several conditions associated with fatally severe accidents were analyzed. The conditions were different for pedestrian accidents and for two-wheeler accidents. Dependencies were also analyzed. In view of the outcomes,
they reasoned that there were dependencies among the factors and that results of the rule-based analysis were reliable with the consequences of other existing investigations utilizing probabilistic models [5] and [6]. Various decision tree algorithms and Naïve Bayes algorithm were combined with various feature selection algorithms and implemented. The performance was studied and the outcome showed that feature selection has enhanced the classification performance of all the models [7]. Association rule mining was used after clustering the accidents based on frequency, using k-means clustering algorithm. Then association rules were obtained which gave the relationship between the factors leading to the accident. Hence, a different approach can be done to obtain the predictability value [8]. Logistic regression model and classification tree model have been compared. The results reveal that the classification tree gives more information when compared to logistic regression model, when evaluated by considering various risk factors [9]. It is deduced that Naïve Bayes classifier works well when the risk factors for a given severity level are independent. Moreover, it also says that decision tree does not need any assumption among dependencies among risk factors, and works well without considering the dependencies among various factors [10].

2. PROPOSED MODEL:

MODULE 1:
It includes collecting the datasets from the records from a particular location. These data can be collected from the police department for various locations. Then these data are cleaned and transformed, where noisy data is reduced and missing values are filled, and the data is stored in a database.

MODULE 2:
It includes the process analysing the given data. First, classification is done using Naïve Bayes classifier. Then, the output of is obtained, which states whether the accident would be critical or non-critical for a given set of attributes. Then, the same data is subject to classification using Decision Tree, to obtain association. Using this data, accidents can be analysed. This system can also be used for predicting the criticality of an accident for a given set of attributes.

3. METHODOLOGY:

3.1. Data Extraction

Large amount of data like accident severity, road crash reports, road traffic accident report, are available in the Transport Ministry and even with private agencies.

3.2. Raw Data Store

The data obtained from various sources are stored in a database.

3.3. Data Pre-processing

Data pre-processing involves Data cleaning. Data transformation which are used to reduce noisy data and to fill missing values. Then the data is stored in database and then searching and analysing can be done using it.

3.4. Data Visualization
Data visualization techniques are used to communicate data or information by displaying it as visual objects (e.g., points, lines or bars). The goal is to communicate information clearly and efficiently to users. The main role of visualization is to depict the accident criticality for each given set of attribute. In the proposed model, the decision tree is given as a representation through which the role of each attribute can be understood.

3.5. Criticality Prediction System

A criticality prediction system is developed, which provides a user interface for the user to interact and enter the value of attributes and to know the criticality of the accident for the given values.

![Diagram of Steps Involved in the System]

**Description of attributes in the data set:**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Group</td>
<td>Describes whether the victim is young, adult or a senior citizen.</td>
</tr>
<tr>
<td>Accident Category</td>
<td>Describes the category of the accident, namely, pedestrian hit, vehicular skid, MVI, divider hit, 2 wheeler, etc.</td>
</tr>
<tr>
<td>Time</td>
<td>The time at which the accident took place.</td>
</tr>
<tr>
<td>Day</td>
<td>The day in which the accident took place, Monday through Sunday.</td>
</tr>
<tr>
<td>Season</td>
<td>The season in which the accident took place, namely, summer, winter, spring and autumn.</td>
</tr>
</tbody>
</table>
4. RESULT AND DISCUSSION:

A classifier is a function used to classify the class variable $Y$ for a given set of variables, $X = (X_1, X_2, \ldots, X_n)$.

These are the feature variables or attribute variables, where $n$ is the number of feature variables.

$f: X \rightarrow Y$

Naïve Bayes and Decision Tree classifiers are used to analyze the factors causing the accident.

4.1. Naïve Bayes Classification

It is based on the Bayes hypothesis. It is easier to design and it is well suited for applications involving large data sets. It works on the Bayes theorem of probability to predict the class of an unknown data set. A Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naïve Bayes model is easier to build and useful for working with large data sets. The given accident data set is first trained, and then a model is created from which prediction can be done. Prediction can be done for required conditions and the criticality of the accident can be predicted. For a class variable $Y$, and feature variables $(X_1, X_2, \ldots, X_n)$, Bayes hypothesis gives the following relation:

$$P(Y|X_1,\ldots,X_n) = P(Y) P(X_1,\ldots,X_n|Y) / P(X_1,\ldots,X_n) \propto P(Y) \prod P(X_i|Y)$$

$$Y = \arg \max = P(Y) \prod P(X_i|Y).$$

4.2. Decision Tree

<table>
<thead>
<tr>
<th>Lighting</th>
<th>Describes the lighting condition at the time of accident, whether it was naturally lit or no light or streetlight.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roadway Feature</td>
<td>Describes the feature of the road whether it was a slope, a curve, an intersection, etc.</td>
</tr>
<tr>
<td>Area around</td>
<td>The area around the accident location, it may be a market, a slope, a curve, an intersection, etc.</td>
</tr>
<tr>
<td>Roadway Type</td>
<td>This attribute describes whether the road is a Highway or a non-highway.</td>
</tr>
<tr>
<td>Accident severity</td>
<td>Describes whether the accident is critical or non-critical. It is the attribute that has to be predicted.</td>
</tr>
</tbody>
</table>
The general motive of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning the decision rules from training data. The decision tree algorithm tries to solve the problem by using a tree representation. Each internal node of the tree corresponds to an attribute while each leaf node represents a class label. Attribute Selection plays a major role in Decision tree. Attribute selection is done by considering factors like Information Gain, Gini Index, etc. The values for each attribute are calculated based on these criteria and stored. The attribute with the high value will be the root of the tree. Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. Hence, an attribute with a lower Gini Index would be preferred.

\[
\text{Gini Index} = 1 - \sum_{i} p_i^2
\]

Classification accuracy is the ratio of correct predictions to the total predictions made. Naïve Bayes and Decision Tree algorithms are applied to the given data set. The algorithms are applied to classify it on the basis of Accident Severity class, where there are two values, namely, Critical and Non-Critical. The class that is predicted depends on the values that are given in the confusion matrix. For a sample information, 800 patterns are given as training data and 200 patterns are given as test data. The value that is obtained is either Critical/Non-Critical. Based upon this, the accuracy of the model is calculated. A confusion matrix is a table that is used to calculate the performance of a prediction model. 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 by using,

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

5. RESULT

The accuracy of each of the models can be calculated individually by using the values of TP, TN, FP and FN from the confusion matrix, which is obtained separately for both the algorithms. The accuracy of Naïve Bayes classifier is found to be more than that of Decision Tree algorithm, and hence, Naïve Bayes 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.

6. REFERENCES:


