International Journal of Engineering Research and General Science Volume 4, Issue 1, January-February, 2016
ISSN 2091-2730

Study the factors affecting road safety using decision tree Algorithms

Naina Mahajan., Dr. Bikram Pal Kaur
Faculty of Information Technology
Punjab Technical University
Punjab, India
nmahajan6@gmail.com, cecm.infotech.bpk@gmail.com

Abstract— The purpose of traffic accident analysis is to find the possible causes of accidents. Road accidents cannot be totally prevented but by suitable traffic engineering and management the accident rate can be reduced to a certain extent. This paper discusses the classification techniques C4.5and ID3 using the WEKA Data mining tool. These techniques use on the NH (National highway) dataset. With the C4.5 and ID3 technique it gives best results and high accuracy with less computation time and error rate.

Keywords— Data Mining, Decision Tree Algorithms, C4.5, ID3, Naive Bayes, WEKA, NH(National highway)

I. INTRODUCTION

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Most commonly used techniques in data mining are: artificial neural networks, genetic algorithms, rule induction, nearest neighbor method and memory based reasoning, logistic regression, discriminate analysis and decision trees. NH provides the efficient mobility and accessibility function. The increasing road accidents have created social problems due to loss of lives and human miseries. Road accidents are essentially caused by interactions of the vehicles, road users and roadway conditions. Each of these basic elements comprises a number of sub elements like pavement characteristics, geometric features, traffic characteristics, road user's behavior, vehicle design, driver's characteristics and environmental aspects. Increase in traffic also brings out extremely severe problem of road accident. The impact of road traffic accident in term of injuries, impairments and fatalities are global social and public health problems. It is now well established that many developing countries face a serious problem of road accidents. Accident fatalities rate in developing countries like India is high in the comparison with those in the developing countries. The population of India has double during the last 30 year while vehicle population has double in the last 5 year. Thus due to increase in the traffic and road accidents certain techniques need to be developed to overcome this problem of accidents. Therefore the factors affecting the traffic accidents have been shared and discussed in this paper.

II. STUDY AREA

The Study has been carried out on the National Highway-1 to reduce the frequency of vehicle accidents passing through NH by using decision tree algorithms.

III. CLASSIFICATION TECHNIQUES

Following are the classification techniques:

C4.5 Technique

Classification algorithms have attracted considerable interest both in the machine learning and in the data mining research areas. Among classification algorithms, the C4.5 system of Quinlan deserves a special mention for several reasons. On the one hand, it represents the result of research in machine learning that traces back to the ID3 system. A decision tree is a tree data structure

consisting of decision nodes and leaves as shown in Figure 1. A leaf species a class value. A decision node species test one of the attributes, which is called the attribute selected at that node.C4.8, implemented in WEKA as J4.8:

Following are the activities of ID3 and C4.5 and also shown in Table 1 and Table 2:

- Permit numeric attributes
- Deal sensibly with missing values
- Pruning to deal with for noisy data

ID3 steps:

- ID3 and C4.5 are algorithms introduced by Quinlan for inducing *Classification Models*, also called *Decision Trees*, from data.
- ID3 works on discrete values only

C4.5 Steps:

- Choose attribute for root node
- Create branch for each value of that attribute
- Split cases according to branches
- Repeat process for each branch until all cases in the branch have the same class

Attributes	Possible values
Age	New, Middle, Old
Competition	Yes, No
Type	Hardware, Software

Table 1: ID3 Uses only Discrete range

Attributes	Possible values
Outlook	Sunny, Overcast, Rain
0Temperature	Continuous
Humidity	Continuous
Windy	True, False

Table 2: C4.5 uses different attribute range

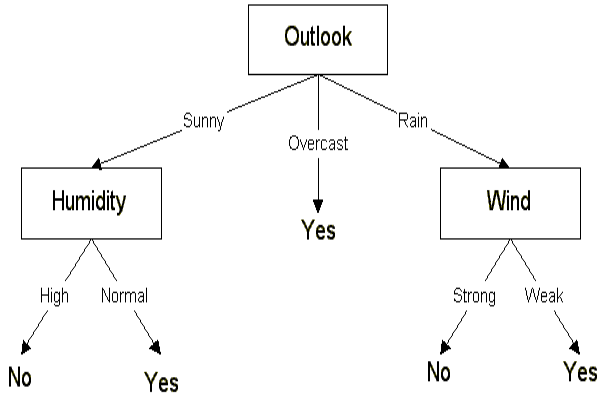

Figure 1: Decision tree

The root node will be that attribute whose gain ratio is maximum. Gain ratio is calculated by the formula.

P(c|x): posterior probability of class (target)

Following example explains the posterior probability:

The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target. Then, transforming the frequency tables to likelihood tables and finally uses the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction as shown in Table 3.

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Table 3: Class with the highest posterior probability

Naïve bayes Classification consider the example given below:

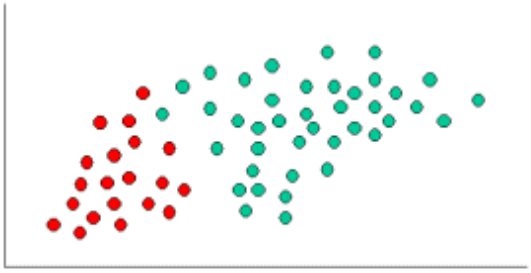

Figure 2: Naive bayes classifier the objects

As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects.

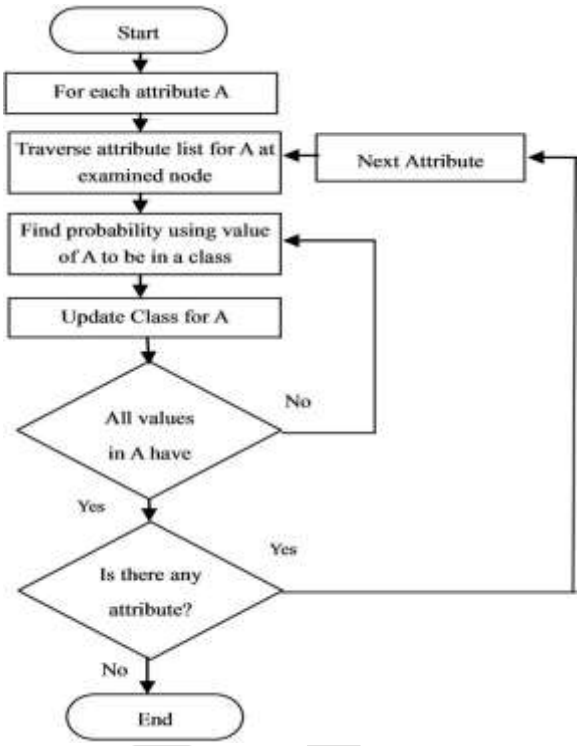

Figure 3: Working of Naive Bayes technique

IV. Experiments

A huge amount of highway accidental data is provided by National Highway Authorities. With this huge scale of data, it is very inefficient or impossible by implementing manual analytical approaches to reach practically meaningful conclusions. In comparing with the conventional statistical methods, decision tree model has shown high efficiency in data analysis and can be used to characterize data and make trend prediction in decision making processes. The decision tree model is therefore chosen to perform the data analysis.

There are major contributing factors for the attributes, which would make the results of decision tree oversize. To get reasonable result, these factors need to be classified into a few groups with losing valuable information. The classified and optimized factors are listed in the following table, which contains 13 distinct groups as shown in Table 4.

1.	**Attention**
2.	**Drink**
3.	**Physical**
4.	**Inexperience**
5.	**Rule**
6.	**Break**
7.	**Mistake**
8.	**Weather**
9.	**Road**
10.	**Sight**
11.	**Age**
12.	**Speed**
13.	**Vehicle**

Table 4: Contributing Factor

V. RESULTS AND DISCUSSIONS

The experiments are conducted through three different aspects with respect to age, season and gender.

7.1 Age group: All the data are categorized into junior, adult and senior according to the drivers' age. By validation, the decision trees generated are tested accurate

7.2 Season: The reasons for accidents in different seasons because of climate conditions. In this section the analysis for two more different groups are presented: winter and non-winter.

7.3 Gender: The experiments were carried out for all the other groups, such as adult, senior, male, female.

VI. EXPERIMENTS RESULTS BY WEKA

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

In this section, the commercial software package, Weka, will be employed for the same data presented in previous section to test the accuracy of program developed in this research. Before the utilization of Weka for the data analysis, two problems need to be fixed first. One is the initial data contains a huge amount of information that would make the results less accurate and hard to understand and analyze. The other is that the format of the original data is not acceptable to Weka.

```
Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0.44 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        2183               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0.0001 %
Root relative squared error              0.0004 %
Total Number of Instances             2183

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0          1         1         1        NO
  1         0          1         1         1        YES

=== Confusion Matrix ===

    a    b   <-- classified as
 1588    0 |    a = NO
    0  595 |    b = YES
```

Table 5: Results through WEKA

VII. CONCLUSION

A series of studies have been carried out to analyze the causes of National highway accidents in using data mining techniques. The objective of this project was to preprocess highway accidental data from Highway Authority for generating human interpretable decision trees and to show the advantage of using decision tree approach for accidental analysis. C4.5decision tree algorithm whose result is compared with ID3 classification techniques. The analysis is conducted through three different aspects with respect to age, season and gender. The Comparison shows good agreement of the result.

REFERENCES:

[1] X-F Zhang, and L. Fan. "A decision tree approach for traffic accident analysis of Saskatchewan highways", *26th Annual Canadian Conference on IEEE*, 5-8 May, 2013

[2] S. B. Kotsiantis, "Decision trees: a recent overview", *Springer Science and Business Media*, Vol. 39, 2011, pp 261-283

[3] Li, L, Zhang, X., "Study of Data Mining Algorithm based on Decision Tree," *International Conference on Computer Design and Applications*, Vol. 1, 2010, pp. 155-158

[4] Rupali Bhardwaj, Sonia Vatta, "Implementation of ID3 Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, 2013, pp 845-851

[5] Deyi Sun, Wing Cheong Lau," Social Relationship Classification based on Interaction Data from Smartphones",IEEE 2nd international workshop on hot topics in pervasive computing, 2013.

[6] M. Mayilvaganan, D. Kalpanadevi, "Comparison of Classification Techniques for predicting the performance of Students Academic Environment", *International Conference on Communication and Network Technologies*, 18-19 December, 2014

[7] Syed Tahir Hijazi, S.M.M Raza Naqvi, "Factors affecting Students Performance: A case of private colleges", *Bangladesh e-journal of Sociology*, Vol. 3, 2006

[8] N. Matthew, G. Sajjan, "Comparative Analysis of Serial Decision Tree Classification Algorithms", *International Journal of Computer Science and Security*,Vol. 3, 2009, pp 230-240

[9] J.R Quinlan, "Induction of Decision Trees Machine Learning*", Kluwer Academic Publishers,* Vol.1, 1986, pp 81-106

[10] Duong Van Hieu, Nawaporn Wisitpongphan, Phayung Meesad, "Analysis of Factors which Impact Facebook Users' Attitudes and Behaviors using Decision Tree Techniques", *11TH International Joint Conference on Computer Science and Software Engineering*, 14-16 May, 2014

[11]Bikram Pal Kaur, Himanshu Aggarwal, " Implementation failures of an Information system: A neuro computing approach", *International journal of computer applications*, Vol. 58, 2012, pp 26-33

[12] Bikram Pal Kaur, Himanshu Aggarwal, "Exploration of success factor of Information system", *International Journal of computer science issues*, Vol. 10, 2013, pp 226-235