Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 11; April-June, 2015 pp. 53-58

http://www.krishisanskriti.org/acsit.html

Enhanced Classification Technique for Traffic Accident Analysis of Highways

Ekta[1] and Rama Rani[2]

[1,2]Department of Computer Science and Engineering DAV University, Jalandhar, Punjab India
E-mail: [1]ekta.uinon@gmail.com, [2]Er.ramaO7@gmail.com

Abstract—*Rapid growth of population in addition to raised financial exercises has supported in gigantic development of engine vehicles. This can be one in all the first factors to blame for road accidents.Accidents may be a unintentional or sudden event.Accidents area unit the results of the failure of individuals, equipment, materials, or setting to react needless to say.Traffic accident leads to loss of life and property.The aim of traffic accident analysis is to search out the attainable causes of accident.Road accidents can't be wholly prevented however by appropriate activity designing and administration the mischance rate may be lessened to a precise degree.This paper introduces the classification techniques C4.5,Naïve bayes and ENHDTA using the WEKA Data mining tool. These techniques use on the NH-1 (Jalandhar to Ambala) dataset. With the ENHDTA technique it gives best results and high accuracy with less computation time and error rate.*

Keywords*: C4.5,NaïveBayes,ENHDTA(Enhanced decision tree algorithm), WEKA datamining Tool,NH(National highway).*

1. INTRODUCTION

Data mining may be a method that uses a range knowledge analysis tools to find patterns and relationships in data which will be accustomed build 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. National highways give the effective versatility and openness capacity. Road accidents area unit primarily caused by interactions of the vehicles, road clients and roadway conditions.Each of these fundamental components embodies various sub components like asphalt qualities, geometric highlights, activity attributes, street client's conduct, vehicle outline, driver's attributes and ecological viewpoints. The effect of street auto collision in term of wounds, disabilities and fatalities are worldwide social and general well beingissues[1]. It is presently entrenched that numerous creating nations confront a significant issue of street mischances. Mishap fatalities rate in creating nations like India is high in the correlation with that in the creating nations.The population of Asian country has double throughout the last thirty year whereas vehicle population has double within the last five year [6]. Decision tree techniques have been generally used to construct classification models [2]There are number of classification techniques for the accidents analysis but EHNDTA technique provides the maximum accuracy and less computation time .The performance of the ENHDTA technique more than the previous classification techniques.

2. STUDY AREA

Study has been completed on the National Highway-1.with the present study, it is conceivable to decrease the recurrence of vehicle mischances going through national expressway NH-1.

3. LITERATURE SURVEY

In 2010,Xue-Fei Zhang etal. [1]utilized the Decision Tree Approach for investigation of accidents in highways. Identifying the real contributing variables for the reasons of accidents and their severity will assist highway safety improvement initiatives by enhanced office configuration and instructive system to deliver the needs because of the progressions in demographics. This paper present an information mining model utilizing ID3 and C4.5 decision tree algorithms to analysis the accidents conducted in highways.

In 2011,S. B. Kotsiantis et al.[2**]**implemented the Decision tree procedures to manufacture characterization models as such models nearly take after human thinking and are straightforward. This paper depicts fundamental decision tree issues and flow exploration focuses. Obviously, a solitary article can't be a complete audit of all algorithms.

In 2011, S.L. Ting et al.[3] the Naïve Bayes content classifier has been broadly utilized because of its effortlessness as a part of both the preparation and characterizing stage. The purpose of this paper is to highlight the execution of utilizing Naïve Bayes in document classification.

In 2012, Mary Slocum et al.[4] examined diverse calculations utilized for creating a decision making (or predictive analysis) system. There are calculations for making decision trees, for example, J48 alongside algorithms for deciding known nearest neighbor (KNN) or grouping when working on pattern

recognition. The objective of this paper is to take a one specific decision tree algorithm called Iterative Dichotomiser 3 (ID3).

In 2012,Rakesh Kumar Singh et al.[5]has studied that with therapid growth of population and expanded monetary exercises has supported in gigantic development of engine vehicles. This is one of the essential elements in charge of street mishaps. It is watched that couple of works have been completed on statistical analysis of mishaps especially on two-path National Highways. In this paper chi-square model is actualized for the investigation reason.

In 2013,KundanMeshram etal.[6] The population is increase day by day and from the beginning of this century the vehicle population is going on increasing. From the past studies they are increased to double within 5 year duration but the length of the road existing is not able to place this much of increasing traffic.This paper introduced the daily estimation reports of accidents.

4. PROBLEM OUTLINE

Traditinal statistical methods for accident analysis take more time and their results are not reliable.But decsion Tree is more effiecient method for characterization of data and apply prediction trends. Data concerned during this paper is extracted from National highway Authority. Probability distribution factor is used to measure the accidents rate and its helps the departments of traffic management, driver training centers andInsurance companies for taking right decisions.

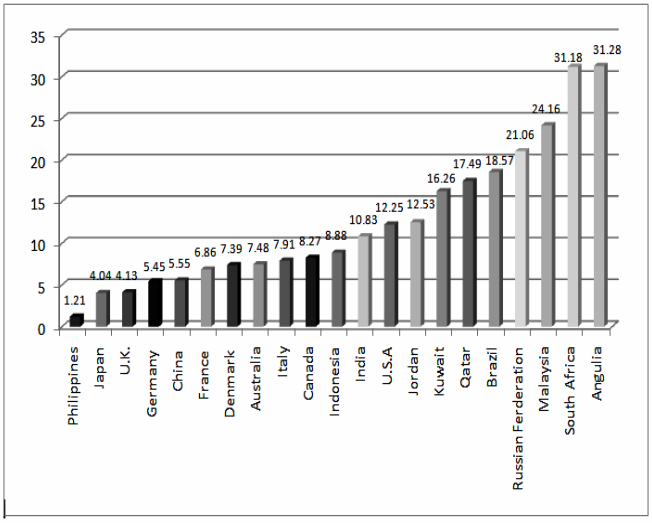

Fig. 1: Accident rates in different countries.

5. CLASSIFICATION TECHNIQUES

5.1 C4.5 Technique

Classification algorithms have attracted extended interest each within the machine learning and within the data processing analysis areas. Among characterization calculations, the C4.5 arrangement of Quinlan merits an exceptional notice for a few reasons[1]. From one perspective, it speaks to the consequence of exploration in machine discovering that follows back to the ID3 framework. [1]. A decision tree is a tree information structure comprisingof decision nodes and leaves. The leaf nodes species the class value. A decision node species a test over one of the attributes, which is called the attributeselected at the node.C4.5 implemented in Weka as J4.8 (Java).

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

ID3 and C4.5 ar algorithms introduced by Quinlan for causation Classification Models, additionally referred to as decision Trees, from data.ID3 works on separate values solely

Table 1: ID3 Uses only Discrete range

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

Table 2: C4.5 uses different attribute range.

Attributes	Possible Values
Outlook	Sunny, overcast, Rain
Temperature	Continuous
Humidity	Continuous
Windy	True, False

Table 3: C4.5 use both numeric and non-numeric training data set.

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Sunny	85	85	False	No
Sunny	85	90	True	No
Overcast	83	78	False	Yes
Rain	70	96	False	Yes
Rain	68	80	False	Yes
Rain	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rain	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rain	71	80	True	No

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

With the highest gain ratio root node will be selected.

Gain ratio can be calculated with the help of this formula.

$$\text{Gain ratio(A)}=\frac{\text{Gain (A)}}{\text{Splitinfo (A)}}$$

The subsequent formula for calculating the entrophy.

$$\text{Entropy(S)} = -\sum_{i=1}^{n} p_i log_2 p_i$$

Entrophy is measure of impurity. The formula for gain attribute is:

$$\text{Gain (A)} = \text{Entropy(S)} - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Entropy } (S_i)$$

5.2 NaiveBayes

The Naivebayes theorem based on classification, estimation and prediction. The bayes algorithm works on large dataset. The algorithm can be easily constructed and uses simple iterative parameters. Naivebayes classifier is fast and effeicient space features. Bayes' theorem with self-rule assumptions between predictors.

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

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

- P(c|x) is posterior probability.
- P(c)is the class prior Probability.
- P(x|c) is the likelihood.
- P(x) Predictor Prior Probability.

Example: From the intial stage we calculates the posterior probability, then create frequency table for every attribute against the target. After that convert the frequency table into likelihood table. At last posterior probability for every class will be calculated. The class with maximum posterior probability is outcome of prediction.

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

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

To understand the Naïve Bayes Classification, we define the example given below:

In this example we classify the twoobjects one in blue(B) color and another one is in red(R) color . Our main target is when the new objects arrive, it should be treated as new case.And determine the object from which class label it relates.when the new object arrive it will be belong to red class because naivebayes theorem based on the past experience.

$$\text{Prior probability (Blue)}=\frac{\text{Number of Blue objects (B)}}{\text{Totalnumber of objects (X)}}$$

$$\text{Prior probability (Red)}=\frac{\text{Number of Red objects (R)}}{\text{Totalnumber of objects (X)}}$$

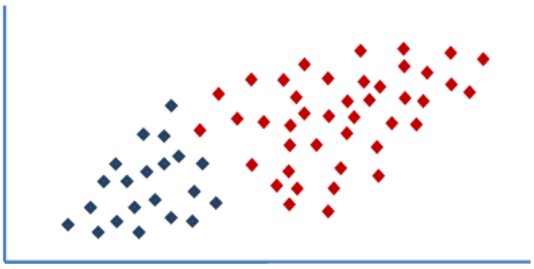

Fig. 2: Naivebayes classifier the objects

Total number of objects(62), objects in blue color(21), objects in red color(41).

Prior probability (Blue)= $21/62$

Prior probability (Red)= $41/62$

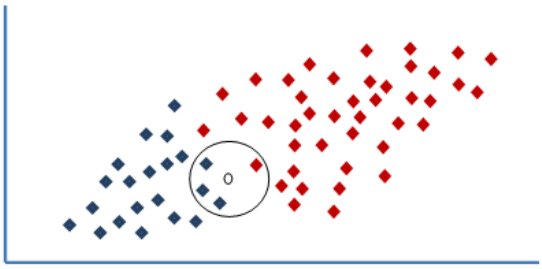

Fig. 3: Naivebayesclassifier definelikehood case.

$$\text{Likelihood of X (B)}=\frac{\text{Number of Blue (obj) near of x}}{\text{Totalnumber of blue cases}}$$

$$\text{Likelihood of X (R)}=\frac{\text{Number of Red (obj) near of x}}{\text{Totalnumber of Red cases}}$$

Probability X (Blue) = $3/21$

Probability X (Red) = $1/41$

Posterior probability of X being Blue=prior probability of Blue*Likelihood of X(B)

$$=\frac{21}{62} \text{ X } \frac{3}{21} =\frac{3}{62}$$

Posterior probability of X being Red =prior probability of Blue* Likelihood of X(B)

$$=\frac{41}{62} \text{ X } \frac{1}{41} =\frac{1}{62}$$

The naïvebayestheoem combine the relationship of both objects i.e blue and red. Bayes theoem is totally based on the prediction. Results are dependent on the past experience.

WORKING PROCESS OF NAÏVEBAYES

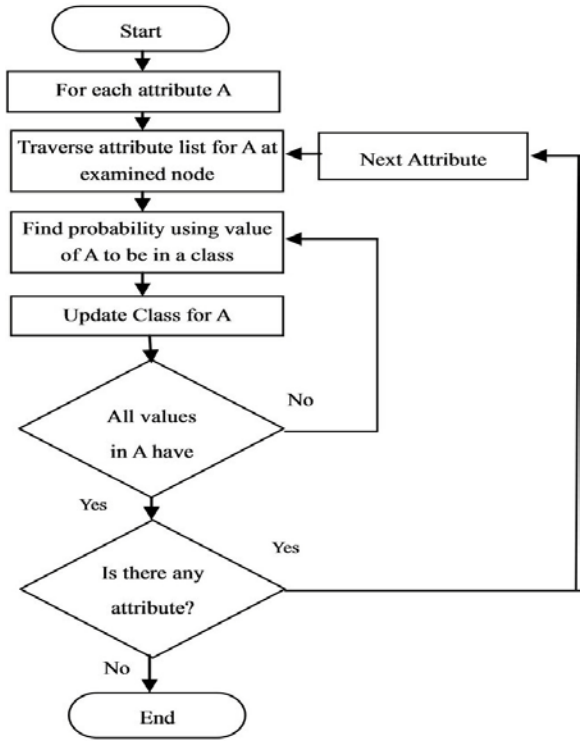

Fig. 4: Working of NaiveBayes technique

5.3 ENHDTA TECHNIQUE

ENHDTA is classification technique that builts a number of step to determine the output of a new data case. ENHDTA technique creates the tree and classify the different arribute values. In this technique, we use the probability factor that is based on the prediction model. ENHDTA is a new approach which provides high reliability and minimize the computational time.

6. EXPERIMENTS

A huge amount of highway accidental data is provided by National Highway Authority, the name of dataset is "Trafficdataset.csv". With this huge scale of data, it is very inefficient or impossible by implementing manual analytical approaches to reach practically meaningful results.

There are major contributing factors for the attributes(MCF), which would make the results of decision tree oversize. To get reliable result, we classify the factors in different groups.

Table 4: Major Contributing Factor

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

7. RESULTS AND DISCUSSIONS

There are three differentaspects with respect to age, season and gender.

7.1 Age group

All the data are categorized into junior, adult and senioraccording 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.

8. EXPERIMENTS RESULTS BY WEKA

Weka may be a assortment of machine learning calculations for data processing tasks. The algorithms will either be applied on to a dataset or known as from your own Java code. The tool contains the knowledge pre-processing, classification, regression, clustering, association rules, and visual image. 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 acceptableto Weka.

Table 5: Results through 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
```

Table6: comparedresults

Error rate

C4.5	Naivebayes	ENHDTA
100	26.77	4

Computation time

C4.5	Naivebayes	ENHDTA
4.52	3.84	0.19

Correctly classified instance

C4.5	Naivebayes	ENHDTA
72	100	100

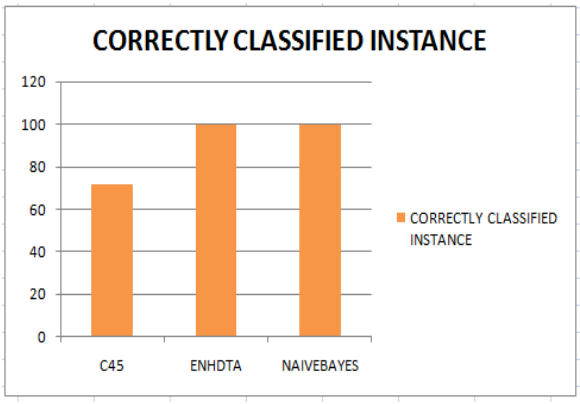

Fig. 5: correctly classified instance

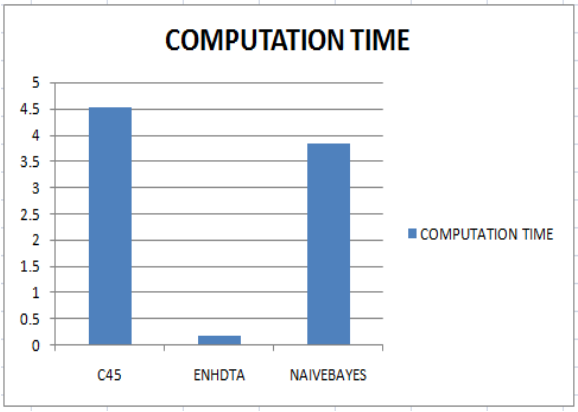

Fig. 6: Computation time

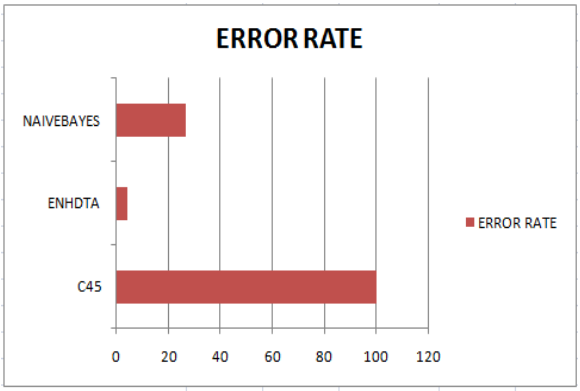

Fig. 7: Error rate

9. CONCLUSION

Progressions of studies have been done to dissect the reasons for National highway mischances in utilizing the data mining methods. The objective of this study is preprocess the information from Highway Authority for producing human interpretable decision trees and to demonstrate the upside of utilizing decision tree approach for incidental examination. ENHDTA decision tree calculates the outcome that is contrasted with C4.5 AND NAVEBAYES arrangement strategies. The investigation is based on three unique aspects concerning age, season and sex. The correlation demonstrates great agreement of the outcomes.

REFERENCES

[1] Xue-Fei Zhang, Lisa Fan,“ A Decision Tree Approach for traffic accident analysis of Saskatchewan highways”,2013 26th IEEE Canadian Conference Of Electrical And Computer Engineering (CCECE)

[2] S. B. Kotsiantis,“ Decision trees: a recent overview”,© Springer Science+Business Media B.V. 2011

[3] Jon Crowcroft, Michael Segal, Liron Levin,“ Improved structures for data collection in static and mobile wireless sensor Networks”,© Springer Science+Business Media New York 2014

[4] RupaliBhardwaj, Sonia Vatta,“ Implementation of ID3 Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering.

[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] Aziz Nasridinov& Sun-Young Ihm& Young-Ho Park,“ A hybrid construction of a decision tree for multimedia contents”, Springer Science+Business Media New York 2013

[7] H. Trevor, T. Robert, and F. Jerome, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction.” Springer, Second Edition, 2009.

[8] N. Matthew, and G. Sajjan, “Comparative Analysis of Serial Decision Tree Classification Algorithms.”, *IJCSS*,vol. 3(3), pp.230-240, 2009.

[9] J.R Quinlan, “Induction of Decision Trees Machine Learning”. Vol.1, pp81-106, 1986.

[10] Lee HY, Lu H, Motoda H (1998) Knowledge discovery and data mining. Knowledge Based Syst 10:401–402