Analysis of Factors of Road Traffic Accidents using Enhanced Decision Tree Algorithm

Naina Mahajan
Department of Information Technology
Punjab Technical University
Punjab, India

Bikram Pal Kaur, PhD
Faculty of Information Technology
Punjab Technical University
Punjab, India

ABSTRACT
Decision tree is a data mining technique which is used to build classification model. ID3 and C4.5 are basic decision tree algorithms that are used to create classification model depending upon the different datasets. In this paper, the enhanced C4.5 algorithm is presented to analyze the traffic accident data using WEKA data mining tool. The main aim of this enhanced approach is to provide a simple, efficient model in classifying the data set. The efficiency of enhanced algorithm is drawn by comparing it with existing algorithms.

Keywords
C4.5, ID3, Decision tree algorithm, NH (National highway), WEKA tool

1. INTRODUCTION
For identification of major cause of accidents, the large amount of data is collected from NHs which is very complex and inefficient to analyze manually. For the analysis, data mining technique can be used to take full advantage of this data set. The results of data mining technique can help highway authority in safety improvements. The traffic accident data analysis can investigate different reasons of traffic accidents. The identification and understanding of these different contributory factors can help public and individual drivers in prevention of major accidents.

In this proposed work, the enhanced C4.5 algorithm is applied on the data collected from NH which is located from Mukerian to Jalandhar, Punjab, India. The existing C4.5 algorithm involves complex calculations which are inefficient when data set is large. The enhanced algorithm can improve computational efficiency and reduce use of memory and error rate. Thus aim of this research is to apply enhanced decision tree algorithm on traffic accident dataset to help highways authority to take decision about training programs for drivers [2]. In this paper decision tree algorithm is used to build classification model due to its significant advantages over the other data mining techniques [3].

2. LITERATURE REVIEW
Following are the authors who have contributed in enhancing decision tree algorithms and analyzing the data using enhanced decision tree algorithms:

Sriram and Yuan [1] proposed enhanced decision tree algorithm to classify human emotions. A customized approach to provide simple and effective prediction model has been proposed. As there are many ways to recognize human emotion like facial recognition, textual conversation etc but emotion detection can also be done using decision tree algorithms. In this paper, the mean and root mean square value is found for all seven emotion values in dataset, in a real time situation. The proposed paper will evaluate the number of correctly and incorrectly classified instances in real time situation.

Zhang and Fan [2] conducted data mining on Saskatchewan Highways traffic data to investigate major contributing factors to traffic collision. C4.5 algorithm is used to create decision tree model. In the proposed paper, the factors affecting the road accidents are used to create decision tree model using enhanced algorithm.

Bedi and Davinder Kaur [3] proposed Review of decision tree data mining algorithms: ID3 and C4.5. The Comparison of ID3 and C4.5 has given. In the proposed paper, the comparison of ID3 and C4.5 is defined on the basis of entropy and information gain.

3. DECISION TREE INDUCTION ALGORITHM
Decision tree learning methods are mostly used in data mining. The goal is to generate a model to predict the value of target variable on the basis of input values [4]. Training dataset is used to generate the tree and test dataset is used to test accuracy of the decision tree. Two major concepts are discussed below [5]:

a. Entropy

Entropy is the level of randomness of data. It is used to find the homogeneity of data attribute. If entropy is zero then sample is homogeneous and if it is one then sample is totally uncertain.

b. Information Gain

Information gain (or can be said as just gain) can be measured as reduction in entropy. The attribute with highest information gain is to be taken as best splitting criterion attribute.

By using these two concepts, the nodes are created and their attributes to split on can be defined. Each leaf node shows the target attribute's value based on input variables shown by path from root to leaf node. First, an attribute that splits data efficiently is chosen as root node in order to generate the small tree. The attribute with higher information is selected as splitting attribute.

The concept of entropy and information gain is used in ID3 and C4.5 algorithms as explained below [6]:

1. ID3 (Iterative Dichotomiser 3)

ID3 algorithm is presented by J.R. Quinlan, 1986, for building the decision tree from the top down. ID3 is a non-incremental algorithm which means that it drives all its classes from a particular set of training instances. The classes generated by ID3 are inductive, that is, given a small set of training instances, the particular classes generated by ID3 are expected to work for all future instances. The distribution of the unknowns must be the same as the test cases. Induction classes cannot be shown to work in every case since they may

classify an infinite number of instances. Note that ID3 may misclassify data. ID3 uses Information gain as splitting criterion. Topmost decision node is the best predictor, it is called root node. The attribute with highest Information Gain is selected as split attribute. Information gain is used to create tree from training instances. This tree is used to classify the test data. When information gain approaches to zero or all instances belong to single target then growing of tree stops. [1].

It develops the tree classifiers in the following three steps:

1. Choose the target attribute and calculate the entropy of attributes.
2. Choose the attribute with highest information gain.
3. Create node which has that attribute. Iteratively apply these steps to new tree branches and stop the growth of the tree after checking the stop criterion.

2. C4.5

C4.5 algorithm is the enhancement to ID3. The number of improvements has been made to ID3 to generate C4.5. C4.5 can take continuous input attribute. C4.5 generate the decision trees from a set of training data in the same way as ID3, using the concept of entropy. The decision trees generated by C4.5 can be used for classification. Because of this reason, C4.5 is also referred to as statistical classifier. J48 is an open source Java implementation of C4.5 algorithm in WEKA tool. While generating a decision tree using C4.5, it needs to deal with the training sets that have records with unknown attribute values by calculating the gain for an attribute by taking only the records where that attribute is given. It follows three steps while the growth of tree:

1. Splitting of categorical attribute is same to ID3 algorithm. The continuous attributes always generate binary splits.
2. Attribute with highest gain ratio is to be chosen.
3. Iteratively apply these steps to new tree branches and stop the growth of the tree after checking of stop criterion. The Information gain bias the attribute with more number of values. C4.5 used a new selection criterion which is Gain ratio which is less biased.

4. IMPLEMENTATION AND RESULTS

In the implementation of ENHANCED ALGO, the traffic data set is collected from location Mukerian to Jalandhar. The factors affecting the traffic accidents which have been considered in this research are followings as shown in Table 1:

Table 1: Factors affecting the traffic accidents

1.	Road Number
2.	Road Type
3.	Speed Limit
4.	Junction Detail
5.	Pedestrian Crossing Human Control
6.	Pedestrian Crossing Physical Facilities
7.	Light Conditions
8.	Weather Conditions
9.	Road Surface Conditions
10.	Special Conditions at Site
11.	Carriageway Hazards
12.	Urban or Rural Area
13.	Did Police Officer Attend Scene of Accident
14.	Accident Location

Here Attention of driver is the target Attribute. The tree generated with enhanced algorithm using the above factors is shown as following in Fig 1:

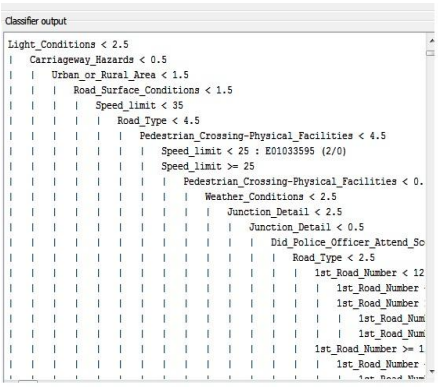

Fig 1: Tree generated with ENHANCED ALGO using WEKA tool

5. CONCLUSION

The enhanced algorithm can get results faster. This enhanced algorithm is applied on traffic data set for analysis of data. The result is compared with C4.5 algorithm. The new enhanced algorithm is computationally efficient when data set is large.

In this paper, enhanced decision tree algorithm is introduced. In this algorithm simplified calculations are used to calculate Gain Ratio. In this way computational efficiency is enhanced and memory usage and error rate is decreased when data set is large.

6. REFERENCES

[1] Sivaraman sriram, xiaobu yuan, "An enhanced approach for classifying emotions using customized decision tree algorithm", *Proceedings of IEEE southeastcon*, January, 2012

[2] 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

[3] Davinder Kaur, Rajeev Bedi, Dr. Sunil Kumar Gupta, "Review of decision tree data mining algorithms: ID3 and C4.5," *Proceedings of international conference on Information Technology and Computer Science*, July 11-12, 2015

[4] Davinder Kaur, Rajeev Bedi, Dr. Sunil Kumar Gupta, "Implementation of enhanced decision tree algorithm on traffic accident analysis", *International Journal of Science Research and Technology*, Vol. 1, 2015, pp. 67-70

[5] Mehdi Mansouri, Mohammad Javed Kargar, " Analysis and Monitoring of the traffic suburban road accidents using data mining techniques", *The Open Transportation Journal*, Vol. 8, 2014, pp. 39-49

[6] 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

[7] 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

[8] 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

[9] Olutayo V.A, Eludire A.A, "Traffic accident analysis using decision trees and neural networks", *International Journal of Information Technology and Computer Science*, Vol. 2, 2014, pp. 22-28

[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