ISSN: 2379-3686 International Journal of Science Research and Technology
Volume 1 Issue 1, p p 67-70 , 15th September 2015

Implementation of Enhanced Decision Tree Algorithm on Traffic Accident Analysis

DavinderKaur[1],Rajeev Bedi[2] ,Dr. Sunil Kumar Gupta[3]
[1]Mtech Student, CSEDept ,BCET Gurdaspur
[2]Assistant Professor,BCETGurdaspur
[3]Associate Professor, BCET, Gurdaspur

Abstract- Decision tree is an important data mining technique which is mostly used to build classification model. ID3 and C4.5 are basic decision tree algorithms that are used to create classification model based on examples. In this paper, we present improved C4.5 algorithms to analyze the traffic collision data. Major contributing factor to traffic collision are identified by applying improved algorithm on traffic data set. Main aim of this improved approach is to provide a simple, efficient model in classifying the given data set. Efficiency of improved algorithm is drawn by comparing it with existing algorithm.

Keywords: - Decision tree, C4.5, information gain, gain ratio.

I. INTRODUCTION

Traffic accident data analysis can investigate different reasons of traffic accident. Identification and understanding of these different contributing factors can help public and individual drivers in prevention of major accidents. For identification of major cause of accident, large amount of data is collected from National highways which are very complex. It is inefficient to carry out data analysis manually. For analysis of data, data mining technique can be used to take full advantage of this data set. Results of data mining technique can help highway authority in safety improvements

Data Mining [5] is a process of inferring information, which is unknown and is useful, form huge amount of data. Data Mining extracts information from large dataset and converts it to an understandable form. Data mining extracts information efficiently than reports and queries. Data mining is part of knowledge discovery process. Classification [4] is a data mining technique that assigns items from dataset to predefined classes by examining the features of new items. Classification technique makes use of training data set and well-defined classes. Training data set consists of classified examples which are used to create model. This model is used to classify unclassified data. Decision tree [2] is a classification model in a tree structure. Decision tree is developed incrementally by breaking down dataset into smaller datasets. Decision tree is generated with decision nodes and leaf nodes. Decision nodes are also known as internal nodes contain splits and splitting attributes. Branches from decision node represents consequences of test on splitting attribute. Each leaf node is associated with a class label. Decision tree is constructed from training set. Then this decision tree is used to classify the tuples with unknown class label. Decision tree algorithms are used to create classification model. Quinlan proposed Id3 and C4.5 decision tree algorithms.

In this proposed work, improved C4.5 algorithm is applied on the data collected from national highway from the location Mukerian to Jalandhar. Existing C4.5 algorithm involve complex calculations which are inefficient when data set is large. The improved algorithm can improve computational efficiency and reduce use of memory and error rate. Thus aim of this study is to apply enhanced decision tree algorithm on traffic accident dataset to help highways authority to take decision about facility design and training programs for drivers. In this paper decision tree algorithm is used to build classification model due to its significant advantages over the other data mining techniques.

II. RELATED WORK

ID3 algorithm is presented by J.R. Quinlan, 1986.ID3 uses Information gain as splitting criterion. Topmost decision node is the best predictor, it is called root node.

In 1993, Quinlan puts forward C4.5 algorithm. C4.5 algorithm is enhancement to ID3.C4.5 can handle continuous input attribute.

Li and Zhang [6] proposed improvement to ID3 Algorithm. They simplify information gain formula.

Hieu and Meesad [11] used decision tree algorithm to classify future behavior and attitude of facebook users. They used C4.5 algorithm to classify data.

Sriram and Yuan [12] proposed enhanced decision tree algorithm to classify human emotions. They proposed a customized approach to provide simple and effective prediction model.

Zhang and Fan [8] conducted data mining on Saskatchewan Highways traffic data to investigate major contributing factors to traffic collision. They used C4.5 algorithm to create decision tree model.

III. DECISION TREE INDUCTION ALGORITHM

Decision tree learning methods are most commonly used in data mining. The goal is create a model to predict value of target variable based on input values. Training dataset is used to create tree and test dataset is used to test accuracy of the decision tree. Each leaf node represents the target attribute's value depend on input variables represented by path by path from root to leaf node.

First, an attribute that splits data efficiently is selected as root node in order to create small tree. The attribute with higher information is selected as splitting attribute [4].

A. *ID3 (Iterative Dichotomiser 3)*

ID3 algorithm is presented by J.R. Quinlan, 1986.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 test data. When information gain approaches to zero or all instances belong to single target then growing of tree stops. [1].

It grows tree classifiers in three steps:

1. Selection of target attribute and calculation of entropy of attributes.
2. Select attribute with highest information gain measure
3. Create node containing that attribute. Iteratively apply these steps to new tree branches and stop growing tree after checking of stop criterion.

The ID3 decision makes use of two concepts when creating a tree from top-down [1]:
1. Entropy
2. Information Gain (as referred to as just gain) Using these two concepts, the nodes to be created and the attributes to split on can be determined.

Entropy

Entropy is degree of randomness of data. It is used to calculate homogeneity of data attribute. If entropy is zero then sample is totally homogeneous and if is one then sample is completely uncertain.

Information Gain

Information gain is decrease in entropy. Attribute with highest information gain is selected as best splitting criterion attribute

$$ET(X, D) = \sum_{j=1}^{k} \frac{|Dj|}{|D|}.ET(Dj)$$

$$IG(X, D) = E(D) - E(X, D)$$

B. *C4.5*

C4.5 algorithm is enhancement to ID3.C4.5 can handle continuous input attribute. It follows three steps during tree growth [3]:

1. Splitting of categorical attribute is same to ID3 algorithm. Continuous attributes always generate binary splits.
2. Attribute with highest gain ratio is selected.
3. Iteratively apply these steps to new tree branches and stop growing tree after checking of stop criterion. 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.

The Gain ratio measure is a selection criterion which is used less biased towards selecting attributes with more number of values [3].

$$GR(X, D) = \frac{IG(X,D)}{SI(X,D)}$$

$$\mathbf{SI}(X, D) = -\sum_{j=1}^{k} \frac{|Dj|}{|D|} \log \frac{|Dj|}{|D|}$$

.

IV. ENHANCED DECISION TREE ALGORITHM

The decision tree induction algorithm C4.5 [3] generates a classification model from training data set. It generates decision tree by calculating information gain of each attribute. The attribute with highest information gain is selected as root. This process repeats to generate tree. Selection of test attribute involves complex calculations and there is repetition of calculations. Thus computational efficiency of tree generation can be affected reduce accuracy of model. In this paper, we proposed a simplified an algorithm which includes simple calculations. This can reduce computational cost of tree generation and enhance accuracy of model.

$ET(D, X) = \sum_{x=1}^{r} \frac{a_x+b_x}{a+b} (D_{1x}+D_{2x})$ where a_x and b_x positive and negative instances

$$\text{Gain Ratio (D, X)} = \frac{Gain(X)}{Split\ Info(X)} = \frac{ET(D)-ET(D,X)}{SI(X)} = \frac{I(a,b)-\left\{\frac{D_1}{N} I(D_{11},D_{12}) + \frac{D_2}{N} I(D_{21},D_{22})\right\}}{SI(D_1,D_2)}$$

$$\left\{\frac{a}{N} log_2 \frac{a}{N} + \frac{b}{N} log_2 \frac{b}{N}\right\} - \left\{\frac{D_1}{N}\left[\frac{D_{11}}{D_1} log_2 \frac{D_{11}}{D_1} + \frac{D_{12}}{D_1} log_2 \frac{D_{12}}{D_1}\right] + \frac{D_2}{N}\left[\frac{D_{21}}{D_2} log_2 \frac{D_{21}}{D_2} + \frac{D_{22}}{D_2} log_2 \frac{D_{22}}{D2}\right]\right\} / \left\{\frac{D_1}{N} log_2 \frac{D_1}{N} + \frac{D_2}{N} log_2 \frac{D_2}{N}\right\}$$

Dividing by $\log_2 e$ and multiplying by N

$$\left\{a \ln \frac{a}{N} + b \ln \frac{b}{N}\right\} - \left\{\left[D_{11} \ln \frac{D_{11}}{D_1} + D_{12} \ln \frac{D_{12}}{D_1}\right] + \left[D_{21} \ln \frac{D_{21}}{D_2} + D_{22} \ln \frac{D_{22}}{D_2}\right]\right\} / \left\{D_1 \ln \frac{D_1}{N} + D_2 \ln \frac{D_2}{N}\right\}$$

As N= a+b

$$\left\{a \ln(1 - \frac{b}{N}) + b \ln(1 - \frac{a}{N})\right\} - \left\{\left[D_{11} \ln(1 - \frac{D_{12}}{D_1}) + D_{12} \ln(1 - \frac{D_{11}}{D_1})\right] + \left[D_{21} \ln(1 - \frac{D_{22}}{D_2}) + D_{22} \ln(1 - \frac{D_{21}}{D_2})\right]\right\} / \left\{D_1 \ln(1 - \frac{D_2}{N}) + D_2 \ln(1 - \frac{D_1}{N})\right\}$$

Gain Ratio (D, X) = $\frac{\frac{ab}{N}-\left\{\left[\frac{D_{11}*D_{12}}{D_1}\right]+\left[\frac{D_{21}*D_{22}}{D_2}\right]\right\}}{\frac{D_1D_2}{N}}$

Above expression involves simple calculation thus calculation complexity decreased. So it reduced memory consumption and decreased error rate.

V. IMPLEMENTATION AND RESULTS

For implementation of EDTALGO, traffic data set is collected from location Mukerian to Jalandhar. This dataset includes 13 attributes. Here Attention is the target Attribute.

1	Drink
2	Physical
3	Experiece
4	Speed
5	Vehical
6	Weather
7	Rule
8	Mistake
9	Road
10	Sight
11	Age
12	Accident
13	Attention

Algorithm is implemented and results are obtained.

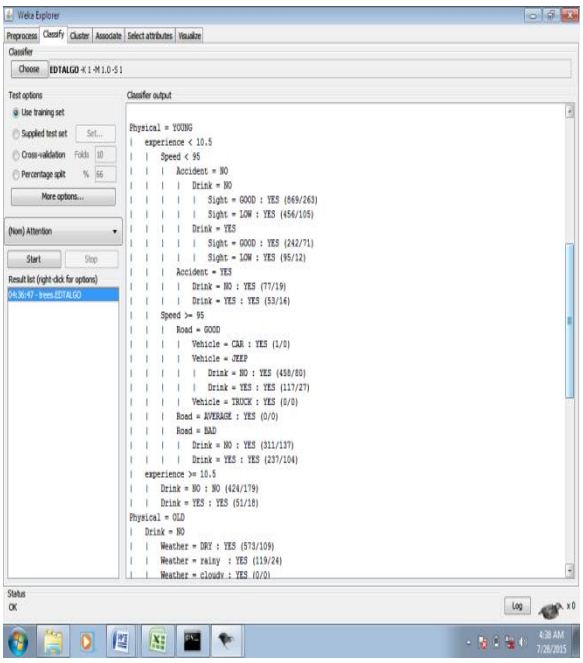

Fig.1 Tree generated with EDTALGO

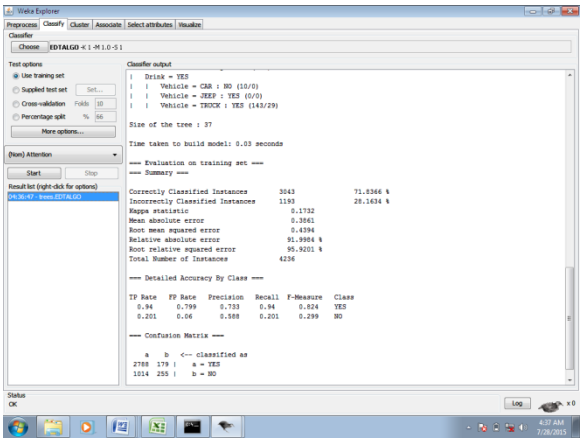

Fig. 2 No of correctly classified instances with EDTALGO

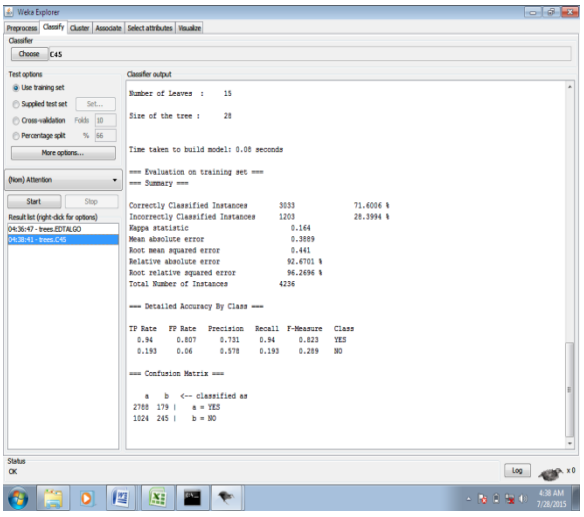

Fig. 2 No. of correctly classified instances with C4.5 algorithm

Results:

	EDTALGO	C4.5
Correctly Classified Instances	71.8366%	71.6006%
Incorrectly Classified Instances	28.1634%	28.3994%

VI. CONCLUSION

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 improved and memory usage and error rate is decreased when data set is large. Improved algorithm can get results faster. This enhanced algorithm is applied on traffic data set for analysis of data. Result is compared with C4.5 algorithm. It is found that incorrectly classified instances decreased with enhanced algorithm. Thus new improved algorithm is computationally efficient when data set is large.

References:

[1]. Fong, P.K. and Weber-Jhanke, J.H (2012), "Privacy Preserving Decision Tree Learning using Unrealized Data Sets", IEEE Transactions on knowledge and Data Engineering, Vol.24,No.2, February 2012, pp. 353-364

[2]. Kabra, R.R. and Bichkar, R.S. (2011),"Performance Prediction of Engineering Students using Decision Tree", International Journal of Computer Applications, Vol.36, No.11, December 2011, pp. 8-12.

[3]. Karaolis, M.A. &Moutiris, J.A (2010), "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees", IEEE Transactions on Information Technology in Biomedicine, Vol.14, No.3, May 2010, pp. 559-566.

[4]. Kesavraj, G. and Sukumaran, S. (2013), "A Study on Classification Technique in Data Mining", 4th ICCNT-2013.

[5]. Sautikar, A.V., Bhujada, V., Bhagat, P.&Khaparde, A.(2014)," A Review paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4,Issue 4, April 2014, pp. 98-101.

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

[7]. Yi-Yang, G. and Man-ping, R. (2009), "Data Mining and Analysis of Our Agriculture based on the Decision Tree", ISECS International Colloquium on Computing, Communication, Control and management, 2009, pp. 134-138.

[8]. Zhang, X.F. and Fan, L.(2013)," A Decision Tree Approach for Traffic accident Analysis of Saskatchewan Highways", 26th IEEE Canadian Conference of Electrical and Computer Engineering(CCECE) 2013.

[9]. Zhang, T., Fulk, G.D. & Tang, W.(2013),"Using Decision Tree to Measure Activities in People with stroke", 35th Annual International Conference of the IEEE EMBS, July 13, pp.6337-6340.

[10]. Suknovic, .M, Delibasic, B., Jovanovic, M., Vukecevic, M., Obradovic, Z.(2011),"Reusable components in decision tree induction algorithm",Comp Stat Februaury 2011.

[11]. Hieu. D.V, Wisitpongphan. N., Meesad.P.(2014),"Analysis of Factors which Impact Facebook Users' Attitudes and Behaviours using Decision Tree Technique", 11th International Joimt Conference on Computer Science and Software Engineering (JcSSE) 2014.

[12]. Sriram. S.,Yuan. X, (2012), "An Enhanced Approach or Classifying Emotions using Customized Decision Tree Algorithm",(2012).