ABSTRACT

Data mining is a process of identification of useful information from large amount of random data. It is used to discover meaningful pattern and rules from data. Classification, clustering, association rules are data mining techniques. Classification is a process of assigning entities to a predefined class by examining the features. Decision tree is a classification technique in which a model is created that anticipates the value of target variable depends on input values. ID3 and C4.5 are commonly used decision tree algorithms. These algorithms are based on Hunt’s algorithm. Goal of this study is to provide review of these decision tree algorithms. At first we present concept of Data Mining, Classification and Decision Tree. Then we present ID3 and C4.5 algorithms and we will make comparison of these two algorithms.

KEYWORDS- Data mining, classification, decision tree

I. INTRODUCTION

Data mining is a process of extraction useful information from large amount of data. It is used to discover meaningful pattern and rules from data. Data mining is a part of wider process called knowledge discovery [4]. The steps of knowledge discovery are

- Selection
- Processing
- Transformation
- Data mining
- Interpretation/Evaluation

Data mining uses two types of approaches i.e supervised learning or unsupervised learning.

A. Classification

Classification is the process of assigning newly presented entities to already defined class by examining the features of entities. Classification is to make decision from unseen cases by building examples of past decisions [2]. There are two steps in classification process.

- In first step, model is built from training data in which value of class label is known. Classification algorithms are used to create model from training data sets.
- In second step, accuracy of model is checked by test data and if correctness of model is satisfactory then the model is used to classify data with unknown class label.

Among classification algorithm, decision tree algorithms is usually used because it is easy to follow and economical to implement.

B. Decision Trees

Decision tree is a classification technique. It is a tree like structure where internal node contains splits and splitting attributes. It represents test on an attribute. Arcs between internal node and its child contain consequences of test. 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 [2].
I. DECISION TREE ALGORITHMS

Decision tree learning methods are most commonly used in data mining. The goal is to 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 depending on input variables represented 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].

Fig. 1 Decision Tree showing whether to go for trip or not depending on weather

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

\[ IG(X, S) = E(S) - E(X, S) \]

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].
\[
GR(X, S) = \frac{I_G(X, S)}{S_0(X, S)}
\]

\[
SI(X, S) = \sum_{j=1}^{k} \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}
\]

Advantages: C4.5 made improvements to ID3 [10]:
1. It can handle both discrete and numerical attributes.
2. It can handle missing value attribute.
3. It can avoid over fitting of decision tree by providing the facility of pre and post pruning.

II. IMPLEMENTATION OF ID3 AND C4.5

![Fig. 3 Import Data Set](image)

![Fig. 4 Define Class Variables](image)

![Fig. 5 Generating Tree based on class attribute services](image)

![Fig. 6 Calculate Error Rate](image)

![Fig. 7 Parameters of C4.5 Algorithm](image)

![Fig. 8 Calculate error rate](image)
IV. CONCLUSION

In this research paper, we presented classification technique decision tree. We presented decision tree algorithm ID3 and C4.5. We focused on key elements of construction of decision tree. We did comparison of ID3 and C4.5 algorithms. It is concluded that C4.5 is more accurate and consume less execution time to mine data with minimum error rate 0.04. C4.5 is a best algorithm for mining a data set.

REFERENCES


