Assignment-01
Decision Tree

Ravi Kumar
14MCMI12
SCIS, University of Hyderabad

March 25, 2015

1 Introduction

Decision tree learning is one of the most successful techniques for supervised classification learning.Decision tree learning is a method for approximating discrete-valued target function, in which the learned function is represented by a decision tree.

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.Each node in the tree specifies a test of some *attribute* of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute.

An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree brach corresponding to value of the attribute .The repeat for the subtree rooted at the new node. An advantage of decision trees is that they easily handle heterogeneous data.

2 ID3 Algorithm

ID3 algorithm, is used to solve the Decision Tree problem and works as follows.

1. Create a root node with the help of *Entropy* and *Gain*.
2. If *Entropy* is zero then
    (a) return the single node with lebel $+ve$
    (b) return the single node with lebel $-ve$
3. otherwise start creating sub tree for each attribute
    (a) A ← the attribute from *Attribute* that *Best* classifies
    (b) the decision tree for Root ← A
    (c) for each possible value v_i, of A,
        i. Add a new tree branch below *Root*, corresponding to the test A = v_i

$$\textbf{Entropy(S)} \equiv \sum_{i=1}^{c} -p_i log_2 p_i$$

$$\textbf{Gain(S,A)} \equiv Entropy(S) - \sum_{v=Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

3 Important Features

1. Strictly Top-down approch.
2. It can only deal with $Nominal$ data.
3. It is not *ronust* in dealing with Noise data.
4. It $ovetfits$ the tree to the training data.
5. It does not handle missing data vlues well.

4 Modules

4.1 Preprocessing

The data is read from a file which is in *txt* format, it contains with the labels which is column name and the data is divided into two. Training(75%) and Testing(25%).

4.2 Best

The *best* attribute is the one with highest $informationgain$.Retun value of from this module is attribute value.

4.3 Gain

This modules returns a gian of attribute with respect to the parse table.Note that parse table may be original table or a filter table.

4.4 Entropy

This modules returns the entropy of a table with respect attribute and attribute value.

4.5 Filter

This module returns the New table according attribute value ,however the columns size will be same but rows size will be decrease.

4.6 Unique

This modules calculates the total number of different attribute value present in particular column.This moudule helps in calculation of *gain* of a attribute.

4.7 Tree Construction

This module creates a Tree which is of the type array(named tree) form.The [0][0] index contains the Root node(Attribute) of the tree.The root index indicate next child node for tree.This process will continue untill it found entropy is zero.

5 Contribution

My contribution to this Assignment is to find out $Filter$ table & implentation of $Treeconstruction.$

6 Conclusion

We implemented ID3 upto root lavel & next lavel of root node.we are still working on this assignment. We didn't perform any testing for completed work.