Modification of information gain measure to select the best group of attributes in a data set for a binary decision tree inducer

N.S. Rathnayaka
Faculty of Engineering, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka
and
J. K. Wijerathna
Faculty of Science, University of Colombo, Colombo 03, Sri Lanka

Introduction

Classification is one of the frequently used techniques in data mining processes which can be applied to accurately predict the target class for each case in a data set. The Decision tree (*DT*) algorithms are one of the powerful classification and prediction methods which facilitate decision making in sequential decision making for a given dataset (Han & Kamber, 2006; Bramer, 2007). The major strengths of the *DT* algorithms are their ability to generate understandable rules, to handle both numerical and categorical attributes and also provide a clear indication of which attributes are most salient for prediction or classification (Kangaiammal, 2013). ID3 and C4.5 are multi splitting algorithms and developed by J. Ross Quinlan in1986 and 1993 respectively. That can be used to Entropy, information gain (*IG*) and Gain ratio as attribute selection measures. These measurements can be utilized to make the binary decision tree to reduce the complexity of the decision tree. If the algorithm identifies more than one attributes with equal *IG* in the data set, then it will select the initial attribute as a splitting node of a tree. This attribute may not be the best attribute for decision making when it is compared with the other attributes of equal *IG*. Therefore, the aim of this study is to improve the *IG* measure to select the best attribute in a dataset and plot a binary decision tree.

Materials and Methods

To overcome the above problem, the relation degree function $AF(A_k)$ was introduced between A_k and C_j ($j=1,2,$ represent two kinds of class values) (C.Jin, L.De-Lin, & M.Fen-xiang, 2009) and the normalization value $V(A_k)$ of the relation degree function was calculated.

$$AF(A_k) = \frac{\sum_{i=1}^{n}(a_{i1}-a_{i2})^2}{n}, \qquad \rightarrow (01) \qquad \text{and} \qquad V(A_k) = \frac{AF(A_k)}{\sum_{k=1}^{l} AF(A_k)}, \qquad \rightarrow (02)$$

Where a_{ij} indicates the i^{th} value of attribute A_k in dataset *D* and j^{th} value of categorical attribute *C* and $0 < k \leq l,$ and *l* is the number of attributes of the dataset *D.*

Then, calculate the values of modified version of *IG*, $MGain(A)$ which is defined as:

$$MGain\ (D, A_k) = (Info(D) - Info_{A_k}(D)) \times V(A_k), \qquad \rightarrow (03)$$

where $Info(D)$ is entropy of *D* and $Info_{A_k}(D)$ (Han & Kamber, 2006) is relative entropy of attribute A_k.

The A_k attribute with highest $MGain(D, A_k)$, is considered as the first splitting attribute of the binary decision tree algorithm. **Algorithm for Binary Decision Tree**

Input: Training Example, *D;* Target Attribute, *C* in; List of Attributes, *l.*
Output: A decision tree.
Handle a case where single attribute is selected

I. Create a root node for the tree
II. **If** tuples in *D* is all of the same class, *C*, then return the single-node tree Root, label with the class *C*.

III. **If** number of predicting attributes is empty, then return the single node tree Root, label with most common value of the target attribute in the examples

Handle the case where multiple attributes have same *IG*

IV. Otherwise

- Find all the possible subsets of each attribute in *D*
- Find the normalization values of relation degree function of each attributes
- Apply Attribute selection method (i.e *IG*) for each subset of each attributes in *D* and then find the subset with maximum impurity value to find the best splitting attribute (*A*).

If there are only one attribute (*A*) with highest impurity value, then

- Label node with the name of attribute *A* and split '*A*' corresponding to two subsets (*D1* and *D2*)

If there is more than one attributes with highest impurity values, then

- Multiply the impurity values by the corresponding *V* values of each attribute and find the corresponding attribute '*A*' with largest value, then
- Label node and split attribute '*A*' into two subsets (*D1* and *D2*)

Stopping Criteria

If tuples in *D1* or *D2* are all of the same class, then

- below this new branch add a leaf node label with class *C* in *D1* or *D2* (step 2)

If *D1* or *D2* is empty, then

- below this new branch add the leaf node with most common value of the target attribute in the examples else below this new branch add the sub tree (attach the node returned by Generate **Binary decision tree**) , End

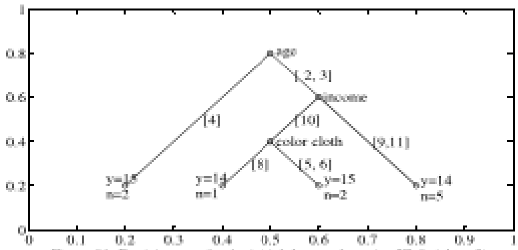
Return root

Results and Discussion

Table 01: Values of conditional entropy, information Gain and modification of IG for possible splitting subsets of each attributes.

Attributes	Possible sets	Conditional Entropy	$Gain(D,A_k)$	V_k	$Gain(D,A_k) \times V_k$
Age	{4} &{2,3}	0.6490	0.3219	0.3810	0.1226
	{3} &{2,4}	0.9245	0.0464		
	{2}&{3,4}	0.9245	0.0464		
Color Cloth	{8}&{5,6,7}	0.8000	0.1709		
	{7}&{5,6,8}	0.6490	0.3219	0.2381	0.0766
	{6}&{5,7,8}	0.9651	0.0058		
	{5}&{6,7,8}	0.9651	0.0058		
	{8,7}&{6,5}	0.9510	0.0200		
	{8,6}&{7,5}	0.8464	0.1245		
	{7,6}&{8,5}	0.8464	0.1245		
Income	{11}&{10,9}	0.9245	0.0464		
	{10}&{11,9}	0.8797	0.0913	0.1906	0.0174
	{9}&{11,10}	0.9651	0.0058		
student	{13}&{12}	0.9245	0.0464	0.1905	0.0088

The Figure 01 and Figure 03 represent the binary decision tree for the initial data set which is used the *IG* and *MIG* attribute selection measure respectively. The Figure 02 and Figure 04 represent the decision tree for the data set which was changed the order of the attributes and applied the same attribute selection measures respectively.



Figuer 01: Decision tree for the initial data set by using IG (height = 3)

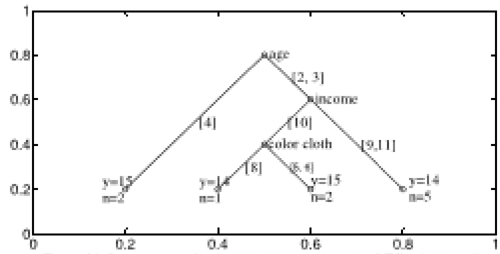


Figure 03: Decision tree for the initial data set by using MIG (height = 3)

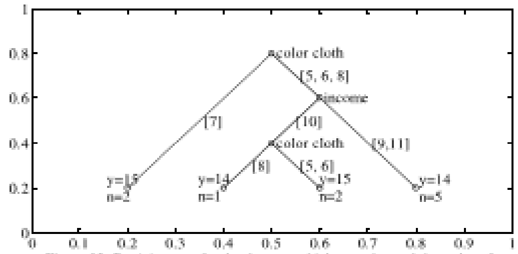


Figure 02: Decision tree for the data set which was changed the order of attributes and used the IG (height = 3)

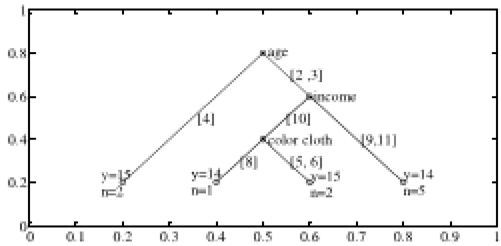


Figure 04: Decision tree for the data set which was changed the order of attributes and used the MIG (height = 3)

Conclusions

The *IG* measure selects the least important attribute to make the decision of the dataset. But, *MIG* measure selects a better attribute than the *IG* measures in decision tree inducer algorithm. When the algorithm classifies categorical data into a binary tree by using *MIG*, it has to find all possible subsets of each attribute, and then the computational efficiency of the algorithm will be deteriorated. Therefore, this algorithm will be more efficient for a data set with a lower number of values of each attribute in a dataset.

References

Bramer, M. (2007). *Principles of Data Mining* (2nd ed.). (I. Mackie, Ed.) London: Springer London Heidelberg New York Dordrecht.

C.Jin, L.De-Lin, & M.Fen-xiang. (2009). An Improvised ID3 Decision Tree Algorithm. *Proceedings of 2009 4th International Conference on Computer Science and Education* , 127-130.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). (M. R. Jim Gray, Ed.) Morgan Kaufmann: Diane Cerra.

Kangaiammal, A. S. (2013). Student Learning Ability Assessment using Rough Set and Data Mining Approaches. *I.J.Modern Education and Computer Science , 5*, 1-11.

Samue, O. D. (2006). *An Exploration of Classification Prediction Techniques in Data Mining:The insurance domain.* Bournemouth,UK: Bournemouth University.