International Journal of Advancements in Research & Technology, Volume 2, Issue 4, April-2013
ISSN 2278-7763

Construction of Decision Tree : Attribute Selection Measures

R. Aruna devi[1], Dr. K. Nirmala[2]

[1]Research Scholar, Manonmanion Sundaranar University & Asst. Professor, Department of Computer Science, Vidhya Sagar Women's College, Chengalpattu, Chennai, Tamil Nadu, India. Email: arunaa_2008@yahoo.com

[2] Associate Professor, Department of Computer Science, Quaid-e- Millath Government College for Women(A),Chennai, Tamil Nadu,India. Email : nimimca@yahoo.com

ABSTRACT

Attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, D, of a class-labeled training tuples into individual classes. It determines how the tuples at a given node are to be split. The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples.This paper, perform a comparative study of two attribute selection measures. The Information gain is used to select the splitting attribute in each node in the tree. The attribute with the highest information gain is chosen as the splitting attribute for the current node. The Gini index measures use binary split for each attribute. The attribute with the minimum gini index as selected as the splitting attribute. The results indicates that predicting a attribute selection in Gini index is more effective and simple compared to Information gain.

Keywords: Heuristics, Information Gain, Gini Index, Attribute selection.

I.INTRODUCTION

Data mining is the extraction of implicit, previously unknown, and potentially useful information from large databases. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form, which is easily comprehensible to humans. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining task can be classified into two categories: Descriptive and Predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make prediction.

II.DECISION TREE

Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The topmost node in a tree is the root node. A tree can be "learned" by splitting the

International Journal of Advancements in Research & Technology, Volume 2, Issue 4, April-2013
ISSN 2278-7763

source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. In data mining, decision trees can be described as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification. The key requirements to do mining with decision trees are: (1) Attribute-value description: object or case must be expressible in terms of a fixed collection of properties or attributes. (2)Predefined classes (target attribute values): The categories to which examples are to be assigned must have been established beforehand (supervised data). (3)Discrete classes: A case does or does not belong to a particular class, and there must be more cases than classes. (4)Sufficient data: Usually hundreds or even thousands of training cases. Decision tree induction is the learning of decision trees from class-labeled training tuples.During tree construction, attribute selection measures are used to select the attributes that partition the tuples into distinct classes.

III.INFORMATION GAIN

This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or "information content" of message. Let node N represents or hold the tuple of partition D. The attribute with the highest information gain is chosen as the splitting attribute for the node N. The expected information needed to classify a tuple in D is given by,

$$\text{Info}(D) = -\sum_{i=1}^{m} pi \log_2 pi$$

Where Pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by $|C_{i,D}| / |D|$. Info(D) is the average amount of information needed to identify the class label of a tuple in D.Info(D) is also known as the entropy of D.The expected information required to classify a tuple from D, based on the partitioning by attribute A is calculated by,

$$\text{Info}_A(D) = \sum_{j=1}^{V} \frac{|Dj|}{|D|} \text{ X Info}(D_j)$$

Information gain is defined as the difference between the original information requirement (i.e. based on the classes) and the new requirement (i.e. obtained after partitioning on A)

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

International Journal of Advancements in Research & Technology, Volume 2, Issue 4, April-2013
ISSN 2278-7763

IV. GINI INDEX

The Gini Index considers a binary split for each attribute. The Gini Index measures the impurity of D, a data partition or set of training tuples as,

$$\text{Gini}(D)=1-\sum_{i=1}^{m} P_i^2$$

Where P_i is the probability that a tuple in D belongs to Class C_i and is estimated by $|C_{i,D}| / |D|$.

When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on A partitions D into D1 and D2, the gini index of D given that partitioning is

$$\text{Gini}_A(D)=\frac{|D1|}{|D|}\text{Gini}(D_1)+\frac{|D2|}{|D|}\text{Gini}(D_2)$$

For each attribute, each of the possible binary split is considered. For a discrete valued attribute, the subset that gives the minimum gini index for that attribute is as its splitting attribute

V. DATASET DESCRIPTION

The main objective of this paper is to select the best attribute measure to construct decision tree.

Table 1

Owns home	Married	Gender	Employed	Class
Yes	Yes	Male	Yes	B
No	No	Female	Yes	A
Yes	Yes	Female	Yes	C
Yes	No	Male	No	B
No	Yes	Female	Yes	C
No	No	Female	Yes	A
No	No	Male	No	B
Yes	No	Female	Yes	A
No	Yes	Female	Yes	C
Yes	Yes	Female	Yes	C

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this paper we wish to select the best attribute measure to construct decision tree. Given the data as in Table 1. The data tuple are described by the attribute owns home, Married, Gender, employed, class.

6.1 INFORMATION GAIN ATTRIBUTE MEASURE

D=10,A=3,B=3,C=4,M=3

Info(D)=$-3/10\log_2(\frac{3}{10})-3/10\log_2(\frac{3}{10})-4/10\log_2(\frac{4}{10})$= 0.521+0.521+0.529= 1.57

We can compute the Attribute "ownshome"

Info $_{\text{ownshome}}$(D)=$5/10[-1/5\log_2(\frac{1}{5})$ - $2/5\log_2(\frac{2}{5})-2/5\log_2(\frac{2}{5})] + 5/10[-2/5\log_2(\frac{2}{5})$ - $1/5\log_2(\frac{1}{5})-2/5\log_2(\frac{2}{5})]$ = 0.761+0.761 = 1.52

Gain(ownshome)=Info(D) - Info $_{\text{ownshome}}$(D)

=1.57-1.52

Similarly we can compute the attributes married, gender, employed.

Table 2

Attribute	Info	Gain
Owns home	1.52	0.05
Married	0.847	0.72
Gender	0.69	0.88
Employed	1.12	0.45

Hence, Gender has the highest information gain among the attribute, so it is selected as the splitting attribute.

International Journal of Advancements in Research & Technology, Volume 2, Issue 4, April-2013
ISSN 2278-7763

6.2 GINI INDEX ATTRIBUTE MEASURE

Total tuples(S)=10
Total Classes(M)=3
Class A=3, Class B=3, Class C=4

Now, compute the gini index for each of the attributes.

Attribute=”ownshome”
Gini(D1)=1-$(1/5)^2$-$(2/5)^2$--$(2/5)^2$ = 0.64
Gini(D2)=1-$(2/5)^2$-$(1/5)^2$--$(2/5)^2$ = 0.64
Gini $_{ownshome}$(D)=5/10(0.64)+5/10(0.64) = 0.64

Similarly we can compute the attributes married, gender, employed.

Table 3

Attribute	GiniIndex
Owns home	0.64
Married	0.40
Gender	0.34
Employed	0.47

Here, Gender has the smallest gini index among the attribute, so it is selected as the splitting attribute.

COMPARISON AND RESULTS

For the comparison of our study, first we used an Information gain as attribute selection measure. Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect. A notable problem occurs when information gain is applied to attributes that can take on a large number of distinct values.

Secondly we used a Gini index as attribute selection measure; it has very time consuming and particularly suitable for multivalue attribute.

VII. CONCLUSION AND FUTURE DEVELOPMENT

In this paper, the comparative study of two attribute selection measure is compared. The Gini index measure is very easy to select the best attribute to construct decision tree because of its simplicity, elegance, and robustness. The results indicate that selection of attribute using gini index is very easy and simple compared to information gain. Possible extension of this work will be developed to use various attribute selection measures like CHAID, C-SEP and MDL- based measures.

REFERENCES

[1]A.K.Pujari, “*Data Mining Techniques*”, University Press, India 2001.
[2]Jiawei Han and Micheline Kamber “Data *Mining Concepts and Techniques*”
[3]S.N.Sivanandam and S.Sumathi, “*Data Mining Concepts Tasks and Techniques*”, Thomson, Business Information India Pvt.Ltd.India 2006
[4] H. Wang, W. Fan, P. Yu, and J. Han.”*Mining concept-drifting data streams using ensemble Classifiers”.*
[5] V. Ganti, J. Gehrke, R.Ramakrishnan, and W.Loh.“*Mining data streams under block evolution”.*
[6] Friedman N, Geiger D, Goldszmidt M (1997) “*Bayesian network classifiers”.*
[7] Jensen F., “*An Introduction to Bayesian Networks”.*
[8] Murthy, “Automatic Construction of Decision Trees from Data”
[9]Website:www.cs.umd.edu/~samir/498/10Algorithms-08.pdf
[10] Website:en.wikipedia.org/wiki/Data_mining