International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9, Issue-7, May 2020

A New Direct Node Data Splitting Technique in Decision Tree Induction

D. Mabuni

Abstract: ***Learning knowledge from the example data and then applying that knowledge on new applications is the goal of machine learning. Split attribute technique is inseparable and an important means of decision tree construction technique. It is a well known fact and universally accepted truth that a difficult and larger data mining model tends to create less significant generalized performance results. Researchers are being continuously trying to find new and the best split attribute techniques during decision tree induction. In this paper a new direct split attribute technique for decision tree induction is proposed based on the mathematical A implies B tautology principle. The experimental results show that this rule is worthy and useful in many real world applications, particularly in medical field. The resulted association relationships are perfectly matching with the expected results.***

Keywords **:** ***A implies B, association relationships, direct split attribute technique, tautology.***

I. INTRODUCTION

Decision tree is a high-flying data classification model predominantly used for data analysis. So far it has been remaining in the top one position of all the data mining algorithms. Decision tree induction principle is dividing and conquers. The strong and desirable important features of decision tree are easy interpretability, scalability, comprehensibility, representation of benchmarking capability and fast convergence property with O(log n) time complexity of its operations. Decision trees are frequently used for both classification and prediction. During the decision tree induction phase the best split attribute method is used for dividing the tuples into optimal homogeneous sub groups. Finding the best splitting attribute is probably the most time taking step during decision tree induction.

Creating an optimal decision tree with a specified complexity is computationally intractable. Generally, in such cases greedy method is advisable to use. Small trees are easy to explain and interpretable. Sometimes tree pruning techniques are very useful.

Node splitting plays a central role in decision tree creation. In this paper an efficient new and direct technique for node splitting is proposed during decision tree induction. This new node splitting technique follows a standard “A implies B” mathematical implication rule along with its frequency counts of satisfying that rule for data comparison between values of attributes or between sets of values of attributes. Machine learning researchers have proposed variety of split measures with different terminologies and futures. There is no standard methodology to compare all these types of node split measures. There is a need of finding and then using a bench mark technique to compare all the node split techniques. Ordering of attribute splits in decision tree induction is very important with respect to running tme, efficiency and effectiveness of the tree.

Evaluation criteria for split selection must consider tradeoffs between accuracy and complexity. Some of the complexity measures are number of leaves, tree height, node size, time complexity etc. The tree accuracy is the average accuracy of all the leaves in the tree. Always try to apply the best split and the best split is one which improves the most possible accuracy. Many splitting measures are interchangeable. Many researchers have proved that final decision tree is insensitive to the any selected split measure.

Revised Manuscript Received on May 30, 2020.
***** Correspondence Author
D. Mabuni*, Department of Computer Science, Dravidian University, Kuppam, India. Email: mabuni.d@gmail.com

II. LITERATURE STUDY

Researchers have been continuously putting their efforts in finding new, simple, standard, and elegant techniques for data processing. In this scenario new techniques are continuously emerging for splitting data in the node of a decision tree model creation process. Some of the well known techniques available for splitting data in the decision tree node are:

1. Information Gain
2. Gain Ratio
3. Gini index
4. Two-ing rule
5. Class attribute mutual information measure
6. Distance function
7. Beta function
8. Chi-squared test
9. Minimum description length

Many node splitting measures are biased towards small sized branches. Accuracy is not dependent on the choice of split measure selection but efficiency, effectiveness and complexity are dependent on split measures. All split attribute techniques are not same but different techniques are useful in different problem contexts. No one split attribute technique is good in all cases of problem handling situations.

Information gain:

Information gain is based on entropy of the dataset before and after the division of the dataset into subgroups. Information gain biases towards smaller partitions with distinct values. Number of splits increases with attributes having larger number of distinct values.

Gain Ratio:

Gain ratio prefers to produce better generalization decision trees. It is a normalized measure and it is considered to be the most reliable and popular splitting technique. Normalized gain and average gain are other variations of this measure.

Gini-Index:

It performs only binary splits on the data. Homogeneity increases as the value of gini-index increases. It biases towards larger partitions and it is computed as one minus sum of squares of probability of each class.

Chi-square:

It is a statistical technique works on categorical values. Difference between the two groups increases as the statistical technique value increases.

There exist many types of split attribute techniques and attribute splitting takes different types of forms. There is no single type of node data splitting technique that will give the best performance in all situations. Some splitting methods are mostly suitable for some datasets and other methods are least suitable for the same datasets.

Mahmood[1] proposed two special algorithms for improving the area under the curve (AUC) value with respect to the standard and popular C4.5 decision tree induction algorithm. B. Chandra and Venkatanaresh babu Kuppili [2] have proposed a new heterogeneous decision tree node split technique for decision tree induction based on the proportionalities of class values in the dividing parts of the dataset. Also authors experimentally verified that proposed method is more accurate, generates lesser height model, and takes less computational time. B. Chandra. et. al. [3] proposed an elegant node data splitting measure for decision tree construction. The new technique works on the principle called distinct classes in the partitions of attribute values. Through experimental results they concluded that the proposed technique creates small and high accurate decision tree classifier.

C. Drummond and R. Holte. [4] it has been mentioned by the authors that in the decision tree induction step node splitting criteria is an important sensitive issue and it must be handled intelligently. David Maxwell Chickering et. al. [5] tried to find small set of potential split points dynamically in the domain of continuous attributes. J. Kent Martin [6] said that the choice of attribute split measure influences very low on the tree accuracy but very high on measures such as efficiency, effectiveness, and pruning. Kweku-Muata Osei-Bryson and Kendall Giles. [7] have studied two types of attribute splitting measures called conditional entropy measures and class attribute information measures. Their experimental results reveal that some datasets are sensitive to the choice of selection of node splitting measures and some other datasets are insensitive to choice of split measure selection.

L. Jiang and C. Li. [8] said that the most popular decision tree node splitting measures are gain; gain ratio, minimum description length and probability estimation techniques. R. Lopez De Mantaras [9] introduced a new distance based measure for attribute splitting in decision tree learning and proved that this distance measure is not biased towards attributes having large number of values. Sebastian Nowozin [10] studied decision tree split measures thoroughly and pointed out that normal measures are biased and those measures must be replaced with corresponding improved measures. Shivaram Kalyanakrishnan and Deepthi Singh [11] have contributed two ideas for improvement of decision tree construction for applications of online advertising and also for categorical attribute values a special cluster based technique is introduced. Yan-Yan Song and Ying LU. [12] studied different and important decision tree construction algorithms with respect to splitting measures, stopping criteria, and pruning etc.

W. M. Gary and T. Ye. [13] have studied and applied more sophisticated cost based model for decision tree creation models. Wray Buntine. [14] said that Mingers studied and compared many decision tree splitting measures and said that randomly selected splitting rule does not decreases the classification accuracy drastically. Xinmeng Zhang and Shengyi Jiang [15] proposed a new maximum similarity metric based split attribute technique useful during decision tree construction step. The goal is choose the best attribute which divides the current data into optimal partitions.

III. PROBLEM DEFINITION

Data splitting in the node during induction of a decision tree is an important feature. None of the existing node splitting techniques is superior in all the situations. Some techniques are good in certain situations and other techniques are good in some other situations. Each technique has its own advantages as well as disadvantages. In this paper an efficient new and direct technique for node splitting during decision tree creation step is proposed. It is very simple and based on a well known “A implies B” standard mathematical rule of implication. This rule is used for finding relationships between the attributes of the training dataset. The relationship between the attributes may be strong or weak. A relationship is said to be strong or weak if it is useful or not. It is not necessary that all strong relationships are useful. In the present study only strong relationships are considered for splitting a node during decision tree model creation.

IV. ALGORITHM

Algorithm Create_New_Split_Decision_Tree(T)
Input:
T root node of the decision tree
Output:
Decision tree model based on new split
1.if current node data size < threshold then
2. convert current node T into leaf node
3. find majority class label of the leaf node
4. return
5.end-if
6.if all the tuples in the current node data belongs to the same class then
7. convert current node T into leaf node
8. find majority class label of the leaf node
9. return
10.end-if
11.find good split attribute using proposed split technique
12.divide data in the current node into partitions using the proposed best split attribute
13.create two new nodes T1 and T2
14.store left partition data into node T1
15.store right partition data into node T2

16.call Create_New_Split_Decision_Tree(T1)
17.call Create_New_Split_Decision_Tree(T2)
18.test the decision tree with test tuples
19.print the tree

A. Proposed Node Splitting Technique

Proposed new and direct decision tree node splitting technique is mainly based on the standard and popular mathematical tautology principle called A implies Y, (A $\Rightarrow$ Y). If A is true then Y always must be true. If A is not true then Y may be either true or false. Y value depends on A only when A is true but not when A is false. The implication rule (Y $\Rightarrow$ A) is not considered in this context. In the present study, efforts are made to find the direct relationship between input attribute and output attribute. That is the association strength of the relationship between predictor and output attributes. More over only strong relationships are considered. The number of true values of A may be more than or equal the number of true values of Y but the number of true values of Y always less than or equal to A. All of the terminologies used in the present study are explained appropriately.

$n(A)$ = number of true values of attribute A
$n(Y)$ = number of true values of attribute Y
$n(A \cap Y)$ = number of common true values of both attributes A and Y. That is number of values of A and Y which are both simultaneously true and associative. This is called association strength between the attributes A and Y. True values of the attribute, A, are compared with the true values of target attribute, Y and taken into consideration only matched values between A and Y with the assumption that the attribute, A, is independent of other attributes.

$$(A \Rightarrow Y) = n(A \cap Y) \quad \ldots\ldots\ldots (1)$$

Proportion or association strength of $(A \Rightarrow Y)$ over n is called true fraction or actual association strength of the rule $(A \Rightarrow Y)$ and it is defined as

$$Proportion\ of\ (A \Rightarrow Y) = \frac{n(A \cap Y)}{n(A)} \quad \ldots\ldots\ldots (2)$$

Aggregate frequency count of the selected implication rule $(A \Rightarrow Y)$ is defined as

$$AFC = \frac{matched\ frquency}{total\ frequency} * Proportion\ of\ (A \Rightarrow Y)$$

$$AFC = Frquency\ ratio * \frac{n(A \cap Y)}{n(A)} \quad \ldots\ldots\ldots (3)$$

Maximum association strength between A and B is 1 and minimum association strength is 0. That is the range of association strength value is between 0 and 1. Association strength value is maximum when $n(A \cap Y) = n(A)$ and minimum when the value $n(A \cap Y) = 0$.

During decision tree construction at each level of the tree creation proposed new splitting measure based on AFC is applied. During induction of the given training dataset AFC score is computed for each attribute then the attribute whose AFC score value is maximum is selected as the best splitting attribute at each level of the decision tree induction. After selecting the best attribute the current data in the current node is split into partitions which increase the height of the tree by one level more. Same process is repeated at each level of the decision tree induction. This newly created decision tree is particularly used for determining the relationships between input and output attributes particularly in the fields of medical, research, defense and so on. Split point measure technique is same but split data is dynamically taken into consideration for partition. In many cases the decision tree is created in depth first search order.

V. EXPERIMENTAL RESULTS

Experiments are conducted by taking one hypothetical dataset and the results show that the proposed new and direct node data splitting technique works correctly according to its intended purpose. The dataset is shown in TABLE-1. The dataset contains five input attributes (exercise, diet control, other bad habits, smoking, drinking) one target attribute (patient has cancer or not), and frequency count of each row. For ease of understanding these attributes are labeled as (A, B, C, D, E Y and FC) respectively. Sixth attribute is the frequency count of occurrences of each tuple instance in the dataset. Frequency count tells the number of patients having been suffered with cancer disease for given set of attribute values.

TABLE-1 Types of attributes of the dataset

S.No	Original Attribute	Rename of the attribute
1	Other bad habits	A
2	Drug addiction	B
3	Diet control	C
4	Smoking	D
5	Drinking	E
6	Affected with cancer	Y
7	Instance frequency	FC

A direct relationship between predictor attributes and target attribute is computed in terms of mathematical numerical measure. This measure shows the direct associate strength between attributes. Associate strength between the attributes increases as the numerical measure value increases. The maximum measure value is 1 and the minimum measure value is 0.

TABLE-2 Hypothetical Cancer Dataset

A	B	C	D	E	Y	FC
0	0	0	0	0	0	0
0	0	0	0	1	0	5
0	0	0	1	0	0	5
0	0	0	1	1	1	10
0	0	1	0	0	0	50
0	0	1	0	1	0	10
0	0	1	1	0	0	10
0	0	1	1	1	1	30
0	1	0	0	0	0	5
0	1	0	0	1	0	10
0	1	0	1	0	0	50
0	1	0	1	1	1	60
0	1	1	0	0	0	10
0	1	1	0	1	0	15
0	1	1	1	0	0	70
0	1	1	1	1	1	60
1	0	0	0	0	0	5
1	0	0	0	1	0	10
1	0	0	1	0	0	10
1	0	0	1	1	1	15
1	0	1	0	0	0	5
1	0	1	0	1	1	15

1	0	1	1	0	0	15
1	0	1	1	1	1	20
1	1	0	0	0	1	10
1	1	0	0	1	0	15
1	1	0	1	0	1	80
1	1	0	1	1	1	70
1	1	1	0	0	0	5
1	1	1	0	1	1	20
1	1	1	1	0	1	80
1	1	1	1	1	1	80

The dataset contains 32 total tuple instances such that each attribute contains16 ones and 16 zeros. At the beginning
Number of zero (0) class labels = 19 and
Number of one (1) class labels = 13

TABLE-3 AFC scores of all input attributes

Attribute true values	A	B	C	D	E
n(Attribute)	16	16	16	16	16
n(Attribute ∩Y)	9	8	7	10	10
Frequency Count of Y	390	460	305	505	380
Frequency count of Attribute	455	640	495	665	445
AFC score	0.482	0.359	0.27	0.475	0.534

First row in the TABLE-3 represents attributes
Second row represents number of 1s in each attribute
Third row represents number of matched 1s in the target attribute
Fourth row represents frequency count of target attribute, Y
Fifth row represents frequency count of input Attribute
Sixth row represents average frequency count (AFC) scores computed through equation (3).

AFC score for the attribute A, AFC (A), is computed as follows

$$AFC = \frac{matched\ frquency}{total\ frequency} * Proportion\ of\ (A \Rightarrow Y)$$

$$AFC = Frquency\ ratio * \frac{n(A \cap Y)}{n(A)}$$

$$AFC\ (A) = \frac{9}{16} * \frac{390}{455} = 0.482$$

Similarly AFC scores of other attributes are computed as

$$AFC\ (B) = \frac{8}{16} * \frac{460}{640} = 0.359$$

$$AFC\ (C) = \frac{7}{16} * \frac{305}{495} = 0.27$$

$$AFC\ (D) = \frac{10}{16} * \frac{505}{665} = 0.475$$

$$AFC\ (E) = \frac{10}{16} * \frac{380}{445} = 0.534$$

The first iteration experimental results are shown in TABLE-2. Out of five input attributes the AFC score of attribute (E) is maximum. So, attribute E is the best splitting attribute and data is divided according to the values of E. The attribute value E = 1 creates its left partition and the value of E = 0 creates its right partition during tree induction. Decision tree created for the given dataset with the proposed technique is shown in FIGURE-1.

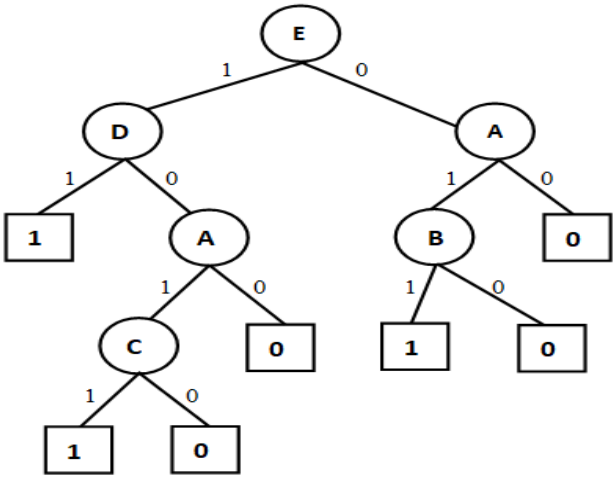

FIGURE-1 Association Relationship Tree

TABLE-4 Attributes and split measures

Attribute	Split measure
A	0.482
B	0.359
C	0.27
D	0.475
E	0.534

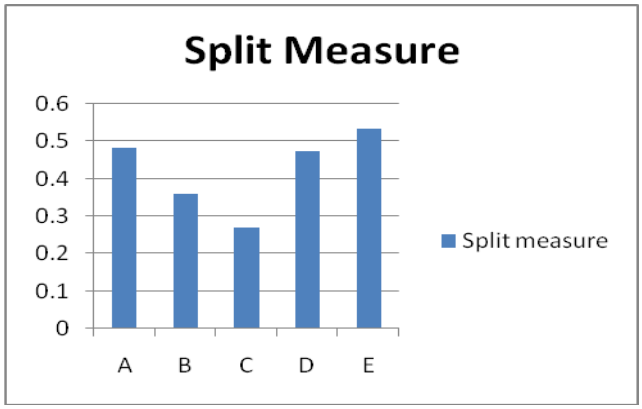

FIGURE-2 Graphical representation of split measures

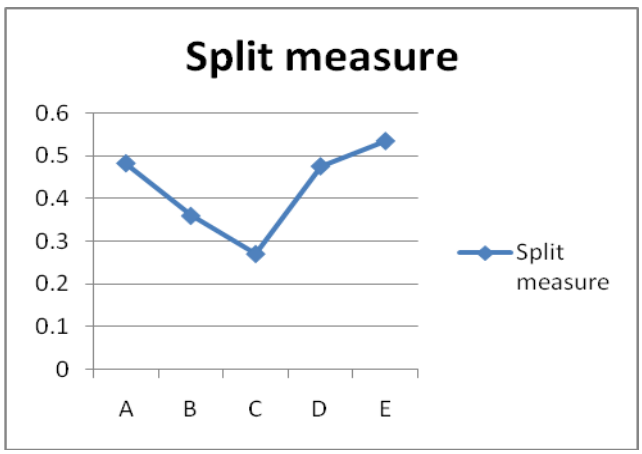

FIGURE-3 Graphical representation of split measures

The tree generated for the dataset is shown in FIGURE-1. In the present paper attribute split measures are normalized for ease of understanding and usage of the relationship strengths. In the FIGURE-1 normalized measure of attribute E is highest and E is selected as the best split attribute.

TABLE-5 Left partition of the root node, E

A	B	C	D	Y	FC
0	0	0	0	0	5
0	0	0	1	1	10
0	0	1	0	0	10
0	0	1	1	1	30
0	1	0	0	0	10
0	1	0	1	1	60
0	1	1	0	0	15
0	1	1	1	1	60
1	0	0	0	0	10
1	0	0	1	1	15
1	0	1	0	1	15
1	0	1	1	1	20
1	1	0	0	0	15
1	1	0	1	1	70
1	1	1	0	1	20
1	1	1	1	1	80

TABLE-6 AFC scores in the second iteration

Attribute	A	B	C	D
n(Attribute)	8	8	8	8
n(Attribute ∩Y)	6	5	6	8
Frequency Count of Y	220	290	225	345
Frequency count of Attribute	245	330	250	345
AFC score	0.673	0.549	0.675	1.0

During second iteration AFC scores are computed using equation – 3. And the results are tabulated in TABLE-5.

$$AFC\ (A) = \frac{6}{8} * \frac{220}{245} = 0.673$$

$$AFC\ (B) = \frac{5}{8} * \frac{290}{330} = 0.549$$

$$AFC\ (C) = \frac{6}{8} * \frac{225}{250} = 0.675$$

$$AFC\ (D) = \frac{8}{8} * \frac{345}{345} = 1.0$$

TABLE-7 Attributes and split measures after E is removed

Attribute	Split measure
A	0.673
B	0.549
C	0.675
D	1.0

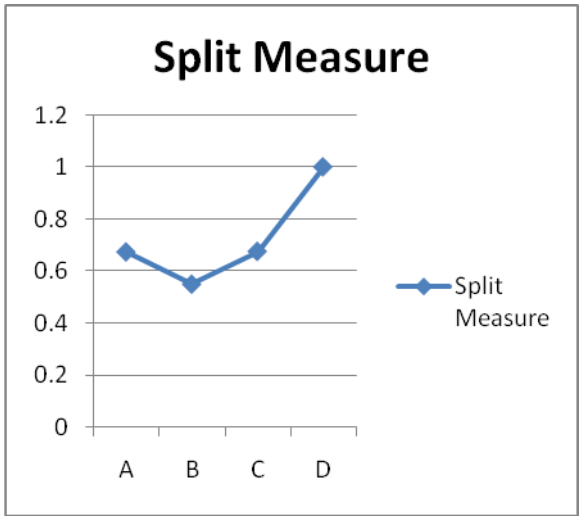

FIGURE-4 Graphical representation of split measures for the left partition of the root node E

The best split attribute is D. So, data is divided according to the values of attribute, D. The tree is grown in the depth first search order and the resulted association relationship tree (ART) is shown in the FIGURE-1

The tree contains six non leaf nodes and seven leaf nodes. Out of seven leaf nodes three leaf nodes represent with 1 class label and the remaining represent 0 class label. One (1) class label means the patient has cancer disease and zero (0) class label means that the patient is not suffering with cancer disease.

Output variable (Y) is a categorical variable which represents either 1 or zero. If Y = 1 means the patient is suffering with cancer disease and if Y = 0 means the patient is not suffering with cancer disease. The relationships shown in the FIGURE-1 explain that there is a strong evidence association relationship between E and Y and because of this reason it was selected as a root attribute of the tree.

If the patient has both the bad habits of drinking (E) and smoking (D) then there is greater chance of suffering with cancer. The attributes E and D are the predominant attributes for getting cancer. If the patient has no habit of drinking then other important factors for getting cancer are drug addiction and other bad habits. The probability of suffering with cancer is more in the left branch than the probability of suffering with cancer in the right branch attribute values of the created decision tree.

TABLE-8 Best split attributes and scores

S.No.	Best Split Attribute	AFC score of Best Split Attribute	Parent Node
1	E	0.534	
2	D	1.0	E
3	A	0.2916	D
4	C	1.0	A
5	A	0.30357	E
6	B	0.728	A

VI. CONCLUSION

A new split attribute technique is proposed for decision tree creation by using the standard mathematical principle called "A implies B" tautology implication. This rule measures the relationships between the attributes in the given dataset. In this study a decision tree is created by assuming that attributes are independent to each other with respect to the output variable. In future the present techniques will be extended for multi class problems and new techniques will be developed for finding combined relationships between the collections of input attributes and the target attribute. Also new efficient and effective attribute splitting techniques will be developed. A relationship between the attributes may be strong or weak depending on the strength of the relationship measure. A strong relationship may be useful or may not be useful. In the future relationship techniques will be thoroughly investigated to find which strong relationships are useful and which strong relationships are useless.

REFERENCES

1. A. M. Mahmood, K. M. Rao, K. K. Reddi, "A Novel Algorithm for Scaling up the Accuracy of Decision Trees.", International Journal on Computer Science and Engineering, vol.2, pp. 126-131, 2010.
2. B. Chandra and Venkatanaresh babu Kuppili, "Heterogeneous node split measure for decision tree construction", **Published in:** 2011 IEEE International Conference on Systems, Man, and Cybernetics, **Date of Conference:** 9-12 Oct. 2011, **Date Added to IEEE *Xplore*: 21** November 2011. **INSPEC Accession Number:** 12387383, **DOI:** 10.1109/ICSMC.2011.6083761
3. B. Chandra, RaviKothari, and PallathPaul, "A new node splitting measure for decision tree construction", ELSEVIER, Pattern Recognition 43 (2010) 2725–2731, Pattern Recognition,
4. C. Drummond, R. Holte. Exploiting the cost (in) sensitivity of decision tree splitting criteria. Proceedings of the Seventeenth International Conference on Machine Learning .pp. 239–246,2000.
5. David Maxwell Chickering, Christopher Meek, and Robert Rounthwaite, "Efficient Determination of Dynamic Split Points in a Decision Tree", Microsoft Research Redmond WA, 98052-6399
6. J. KENT MARTIN, "An Exact Probability Metric for Decision TreeSplitting and Stopping", Machine Learning, 28, 257–291 (1997)1997 Kluwer Academic Publishers. Manufactured in The Netherlands.
7. Kweku-Muata Osei-Bryson and Kendall Giles, "Splitting methods for decision tree induction: An exploration of the relative performance of two entropy-based families", *Information Systems Frontiers* volume 8, pages195–209(2006), Published: July 2006, Springer
8. L. Jiang, C. Li, "An Empirical Study on Class Probability Estimates in Decision Tree Learning," journal of software,vol.6, pp.1368-1372,2011.
9. R. Lopez De Mantaras, "A distance based attribute selection measure for decision tree induction", Machine Learning, 6, 81-92 (1991) © 1991 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.
10. Sebastian Nowozin, "Improved Information Gain Estimates for Decision Tree Induction", Appearing in Proceedings of the 29 th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012. Copyright 2012 by the author(s)/owner(s).
11. Shivaram Kalyanakrishnan and Deepthi Singh, Ravi Kant, "On Building Decision Trees from Large-scale Data in Applications of On-line Advertising", In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 669--678, ACM, 2014, CIKM'14, November 3–7, 2014, Shanghai, China. Copyright 2014 ACM 978-1-4503-2598- 1/14/11 ...$15.00. http://dx.doi.org/10.1145/2661829.2662044.
12. Yan-Yan Song and Ying LU, "Decision tree methods: Application for classification and prediction", Shanghai Archives of Psychaitry 2015 15 April 25; 27(2) 130-135, pmcid: pmc4466856
13. W. M. Gary, T. Ye. Maximizing classifier utility when there are data acquisition and modeling costs. Data Min Knowl Disc, vol.17,pp.253–282,2008.
14. Wray Buntine, "A Further Comparison of Splitting Rules for Decision-Tree Induction", Article **(PDF Available)** *in* Machine Learning 8(1):75-85 · January 1992 *with* 615 Reads, DOI: 10.1007/BF00994006 · Source: DBLP
15. Xinmeng Zhang and Shengyi Jiang, "A Splitting Criteria Based on Similarity in Decision Tree Learning", JOURNAL OF SOFTWARE, VOL. 7, NO. 8, AUGUST 2012, © 2012 ACADEMY PUBLISHER doi:10.4304/jsw.7.8.1775-1782

AUTHORS PROFILE

D. Mabuni, completed M.Sc. (Computer Science), MCA and M.Phil. (Computer Science). Currently working as Assistant Professor in the Department of Computer Science at Dravidian University, Kuppam, Andhra Pradesh, India. My interested research areas are Data Mining, Databases, and User Interfaces.