Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Decision Tree Based Approaches in Data Mining

Shahrukh Teli
M-Tech Student, MPSTME,
SVKM'S NMIMS University,
Mumbai, India

Prashasti Kanikar
Assistant Professor, MPSTME,
SVKM'SNMIMS University,
Mumbai, India

Abstract—Decision tree learning is the most popular and powerful approach in knowledge discovery as well as in data mining. This is used for exploring large and complex bodies of data in order to discover useful patterns. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Classification algorithm processes a training set containing a set of attributes. ID3 algorithm is the most widely used decision tree based algorithms. The main disadvantage of the ID3 algorithm is that it chooses the attribute based on occurrence not on the importance. So in this paper we are going to discuss the ID3 based algorithms which select the attribute based on the importance. This paper discusses about the various ID3 based decision tree algorithms such as Improved ID3 based on attribute importance include the Association function, Attribute importance and Attribute weight. Finally compared all the above algorithms based on the working strategies, characteristics, and features.

Keywords— Decision Tree learning, Classification, ID3, Improved ID3, Entropy

I. INTRODUCTION

Data mining concerns theories, methodologies, and in particular, computer systems for knowledge extraction or mining from large amounts of data. Data mining is a method to extract the knowledge and information from a large number data such as incomplete, noisy and random. In these data, the knowledge and information are implicit, this information do not know in advance, but it is useful. In today's world classification is an important technique in data mining. For classification the data use decision tree. The decision tree is important tool in data mining to do. Compare with the other, decision tree is a faster and more accurate. As a very important and widely used technology in data mining, data classification is currently used in many fields. The purpose of data classification is to construct a classification model, which can be mapped to a particular subclass through the data list in the databank. The decision tree algorithm is a more general data classification function approximation algorithm based on machine learning [2].

The various algorithms which are used for the classification of data are decision trees, linear programming, neural network and statistics. Among these algorithms Decision trees is one of the most popular and powerful approaches in data mining. The science and technology of exploring large and complex data to discover useful patterns this area is most importance for modeling and knowledge extraction from the data which are available. Both theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate. Decision trees, originally implemented in decision theory and statistics. The benefits of decision tree in data mining 1) It able to handle variety of input data such as nominal, numeric and textual. 2) It process the dataset that contain the errors and missing values. 3) It is available in in varies packages of data mining and number of platform [9].

A decision tree is a tree structure which classifies an input sample into one of its possible classes. Decision trees are used to extract knowledge by making decision rules from the largeamount of available information. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new data. This paper reviews different algorithms to classify the data using decision tree. The following example illustrates working of decision tree algorithm[10].

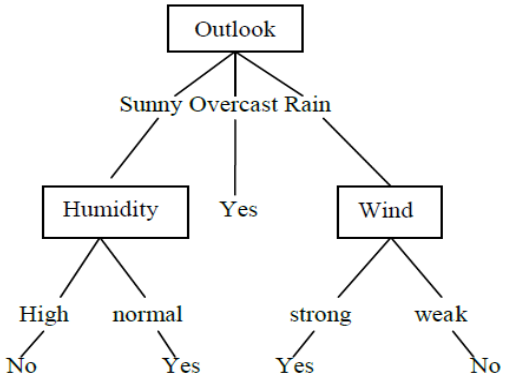

Figure 1: Decision Tree

The rest of the paper is organized as follows. Section 2 provides a working of the decision tree algorithm. Section 3 explain about the types of various ID3 based algorithms. Section 4 describe about the Comparative study of various ID3 based Algorithms for generating the tree and rules. And finally conclude the paper in Section 5.

II. RELATED WORK

In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain. Iterative Dichotomiser 3 is an algorithm used to generate a decision tree [15].

ID3 is the most important decision tree generation algorithm, which is proposed in 1986 by Quinlan. But ID3 has main disadvantage is select the attribute which as many values. To solve this problem researches discuss the various approaches.

ID3 is a supervised learning algorithm, based on information entropy. It is developed from a set of datasets from several classes. The algorithm proposed a set of rules that allows it to predict the class of an item. ID3 identifies the attribute of the class that differentiate from others class. ID3 know the all dataset value in advance. To determine which attribute of the training sample set are important and need to be included atwhich position of the decision tree, ID3 uses the concept of entropy and information gain.Entropy(S) is measures the how random the class distribution is in S. and Information gain measures the how given attribute classify the training examples to their target classification [16].

The ID3 algorithm has dataset and S as the root node. On each step of the algorithm, it going through every unused attribute of the set S and calculates the entropy I(S) (or information gain IG(A)) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then classify based on selected attribute (e.g. age < 50, 50 <= age < 100, age >= 100) to produce subsets of the data. The algorithm continue for each attribute but select the attribute is does not select before [1].

The ID3 algorithm uses information gain to decide the splitting attribute. Given a collection S of possible outcomes, Entropy is defined as it measures the uncertainty present in data set is calculate by equation 1 [13] [17].

$$\text{Entropy}(S) = -\sum p(x) \log_2 p(x) \quad (1)$$

Where,

S is dataset for which entropy is calculated

X is set of classes in dataset

P(x) is proportion/probability of the number of element in class x to the number of element in the set S

When I(S) =0 then the dataset perfectly classify i.e. all element in S are of the same class.

The information gain can be defined as it measure the difference between total entropy to the after the set is split on an attribute. In other words how much uncertainty in S is reduced after splitting set S on attribute A.?

$$IG(S) = I(S) - \sum p(t) * I(t)$$

Where,

I (S) is entropy of dataset

T is subset created from splitting S by attribute A

P (t) is proportion/probability of the number of element in class t to the number of element in the set S

I (t) is entropy of subset t

The basic algorithm steps are as follows

Input: Data partition, S, which is a set of training tuples and their associated class labels; attribute list, the set of candidate attributes; Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset [2][7].

Output: A decision tree.

The main steps of ID3 algorithm is [14]:

- Select the all attributes from different levels of decision tree nodes.
- Calculate the information gain for each and every attribute.
- Use the information gain as the attribute selection criteria/measures and select the attribute with the largest information gain to decide it the root node of the decision tree.
- Branches of the decision tree are calculate by the different information gain values of the nodes.
- Build the decision tree nodes and branches recursively until a particular dataset of the instances belongs to the same group.

Disadvantages of ID3 Algorithm:

Selecting information gain as the attribute selection criteria by ID3 algorithm has the problem of attribute bias. That is this method has the problem of inclining to choose attributions which have more values. While it is not suitable to select the attributions having more values which are not give the best attributes. So few major disadvantages of ID3 Algorithm are mentioned below:

- It chooses the attributes based on their occurrence rather than their importance.
- It does not deal with noisy data sets properly.
- It over-fits the tree to the training data.
- It creates complex trees without pruning unnecessarily.

III. TYPES OF ID3 BASED ALGORITHMS

The main problem of ID3 algorithm is to select the attribute with many values among all the attributes in the dataset. To overcome this problem various researchers have suggested different modified approaches like improved ID3 based on attribute importance, attribute weight and combining ID3 with association function. These algorithms are analyzed and discussed below.

1. Improved ID3 based on attribute importance include the association function[3]

The Association Function based ID3 algorithm based on attribute importance of each attribute. The information gain is to be combine with attribute importance. It form new selection of attribute method to construct decision tree. It uses correlation function method (AF). This method not only overcome the problem of ID3 i.e. take value with more values but it also represent the relation between the attribute and all elements. So the obtained relation degree value of attribute and reflect its importance.

AF algorithm: This algorithm considersthe A is attribute of data set S, category attribute of dataset S is C. so the relation degree function between A and C is represents as difference between the two cases and summation of this cases at n times i.e. total number of attribute. And normalization of AF is follows. The ratio of AF (K) to the total number of AF (m). Where, m is number of attribute and stored in R. the calculation of gain is done by using

$$\text{Gain } (A) = I (S_1, S_2, S_3 \ldots S_m) - E (A) * V (A)$$

Form this equation compute the gain for each attribute and select the root attribute based on highest gain among all attribute. By this way it overcome the problem of ID3 algorithm.

Advantages:

- Selection of attribute with more values is not going to be selected Disadvantages:
- Computational complexity is high

2. Improved ID3 based on attribute importance[4]

Thisalgorithm works based on attribute-importance. The improved ID3 algorithm uses attribute-importance to increase information gain of attribution which has fewer attributions. The experimental analysis of the data shows that the improved ID3 algorithm can get more reasonable and more effective rules. The improved algorithm through introducing attribute importance the attributes with fewer values and higher importance, the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributions with more values[11].

3. Improved ID3 algorithm based on attribute importance and attribute weight[6]

This algorithm mainly focus on the calculation of gain. Because the ID3 algorithmtakes more computation time to compute the gain. So for solving this problem it usesTaylor formula to transform the complex log operation into addition and multiplication operation to reduce the computational time and for the attribute importance it uses attribute similarity. In attribute similarity it calculates the similarity present between the decision attribute and condition attribute. Based on the similarity compute the ratio of similarity of single attribute to the sum of all attribute similarity. This ratio use as weight of that attribute. If the higher the weight then it has highest priority of those attribute. So the gain computed by using the following formula

$$\text{Gain } (A) = I' (S1, S2, S3 \ldots Sm) - E(A) * W(A)$$

Where, I is entropy, W is weight of an attribute.

Advantages:

- It reduces the computational complexity
- It also overcomes the problem of selection of the attributes with more values

IV. COMPARATIVE STUDY OF THE ALGORITHMS

This section comprises all the algorithms based on some parameters and they are given below.

TABLE I. COMPARISON OF ALL ALGORITHM

Name of the Algorithms	**ID3 Algorithm**	**Improved ID3 based on attribute importance include the association function**	**Improved ID3 based on attribute importance**	**Improved ID3 algorithm based on attribute importance and weight**
Working of an algorithm	• Select the attributes from different levels of decision tree	• Select the attributes • Calculate the information gain is to be combine with	• Select the attributes • Calculate the information gain	• Select the attributes • Calculate the information gain is to be combine with

	nodes • Calculate the information gain for each and every attribute • Select the attribute with the largest information gain to decide it the root node of the decision tree	the association function. • It form new selection of attribute method to construct decision tree	is to be combine with attribute importance. • Using this gain construct decision tree	Taylor formula and attribute similarity • It form new selection of attribute method to construct decision tree
Characteristics of the algorithm	• It uses a greedy approach by selecting the best attribute to split the dataset on each iteration	• It uses the association function to increases the attribute importance	• It uses attribute-importance to increase information gain of attribution which has fewer attribution	• To transform the complex log operation into addition and multiplication operation use Taylor formula
Feature is use in the classification	• It use the information gain and entropy	• It use the information gain as well as attribute importance based on association function and entropy	• It uses the attribute importance	• It uses attribute importance and attribute weight
Advantages	• The rules is obtained from dataset is understandable • The tree is build is fastest and simple • Whole dataset is searched to create tree	• Selection based on attribute importance not attribute as more values	• solve the problem of inclining to choose attributions with more values • It give more effective classification rules	• Computational complexity is less as well as It also overcomes the problem of selection of the attributes with more values
Disadvantages	• The calculation method based on information entropy is to choose attributes which have many values	• Computational complexity is high	-	• Computational complexity is less but it uses calculation of weight based on attribute similarity

Table I shows comparison of various existing ID3 algorithm with modified ID3 algorithms based on working, characteristics, features, advantage and disadvantage. From above comparison table Improved ID3 algorithm based on attribute importance and weight is best approach.

V. CONCLUSION

A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. In this paperdiscussed the various algorithms that are used for classification of data using Decision tree algorithm like ID3, Improved ID3 based on attribute importance, based on weight, based on attribute importance and weight which are used for generating the tree. Compared thealgorithms based on the working of algorithm, characteristics of an algorithm, feature of an algorithm, advantages and disadvantages.

ACKNOWLEDGMENT

The research was supported by faculty guide Dr. Dhirendra Mishra. I am grateful to them for sharing their pearls of wisdom with me during the course of research.

REFERENCES

[1] B M, Patil. "Performance Analysis on Uncertain Data using Decision Tree."International Journal of Computer Applications 96, no. 7 (2014)

[2] Changala, Ravindra, Annapurna Gummadi, G. Yedukondalu, and U. N. P. G. Raju. "Classification by decision tree induction algorithm to learn decision trees from the classlabeled training tuples." International Journal of Advanced Research in Computer Science and Software Engineering 2, no. 4 (2012): 427-434.

[3] Jin, Chen, Luo De-lin, and Mu Fen-xiang. "An improved ID3 decision tree algorithm." In Computer Science & Education, 2009. ICCSE'09. 4th International Conference on, pp. 127-130. IEEE, 2009.

[4] Yuxun, Liu, and Xie Niuniu. "Improved ID3 algorithm." In Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on, vol. 8, pp. 465-468. IEEE, 2010.

[5] Chen, Xiao Juan, Zhi Gang Zhang, and Yue Tong. "An Improved ID3 Decision Tree Algorithm." In Advanced Materials Research, vol. 962, pp. 2842-2847. 2014.

[6] Luo, Hongwu, Yongjie Chen, and Wendong Zhang. "An Improved ID3 Algorithm Based on Attribute Importance-Weighted." In 2010 2nd International Workshop on Database Technology and Applications, pp. 1-4. 2010.

[7] Rui-Min, Chai, and Wang Miao. "A more efficient classification scheme for ID3." In Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, vol. 1, pp. V1-329. IEEE, 2010.

[8] Chahal, Hemlata. "ID3 Modification and Implementation in Data Mining."International Journal of Computer Applications 80, no. 7 (2013): 16-23.

[9] Li, Linna, and Xuemin Zhang. "Study of data mining algorithm based on decision tree." In Computer Design and Applications (ICCDA), 2010 International Conference on, vol. 1, pp. V1-155. IEEE, 2010.

[10] Chourasia, Shikha. "Survey paper on improved methods of ID3 decision tree classification." International Journal of Scientific and Research Publications(2013): 1-4.

[11] Bhagwatkar, Priti, and Parmalik Kumar. "Improved the Classification Ratio of ID3 Algorithm Using Attribute Correlation and Genetic Algorithm." International Journal of Advanced Computer Engineering and Communication Technology (IJACECT) ISSN (Print): 2319-2526, Volume-3, Issue-2, 2014

[12] Rongtao, Ding, Ji Xinhao, Zhu Linting, and Ren Wei. "Study of the Learning Model Based on Improved ID3 Algorithm." In Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on, pp. 391-395. IEEE, 2008.

[13] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1, no. 1 (1986): 81-106.

[14] Bahety, Anand. "Extension and Evaluation of ID3–Decision Tree Algorithm."Entropy (S) 2: 1.

[15] Gao, Wushi, Yunfeng Dong, and Kan Li. "The Research and Application of Improved Decision Tree Algorithm in University Performance Analysis." InProceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Atlantis Press, 2013.

[16] Quinlan, J. Ross. "Simplifying decision trees." International Journal of Human-Computer Studies 51, no. 2 (1999): 497-510.

[17] ID3 algorithm "http://en.wikipedia.org/wiki/ID3_algorithm"

