International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 02 (February 2015), PP.44-47

Implementation of Improved ID3 Algorithm to Obtain more Optimal Decision Tree.

Prof. Prashant G. Ahire, [Student]Sanket Kolhe, Kunal Kirange,
Hemant Karale, Abhilasha Bhole
Assistant Professor, Department of Computer Engineering Pimpri Chinchwad
College of Engineering University of Pune

Abstract:- Decision tree learning is the discipline to create a predictive model to map the different items in the set and respective target values and associate them in a way that is true to every element. This concept is used in statistics, data mining and machine learning due to its simple and effectiveness.

Among the various strategies available to construct the decision trees ID3 is one of the simplest and most widely used decision tree algorithm, but ID3 algorithm gives more importance to attributes having multiple values while selecting node. This major shortcoming affects the accuracy of decision tree. In this paper we are proposing improvement in ID3 algorithm using association function (AF). The Experimental result shows improved ID3 algorithm can overcome shortcomings of ID3 which will also improve the accuracy of ID3 algorithm.

Keywords:- Algorithm, ID3, association function.

I. INTRODUCTION

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large.

One of the most popular classification techniques is decision tree induction. Decision tree is an abstraction to visually and explicitly represent the dependencies between attributes of elements in a set and their sorting on the tree determines a set of rules which explain and summarize the relations of all items throughout the set.

One of most important algorithm for generation of decision tree is ID3 which was developed by Professor J. Ross Quinlan.ID3 constructs decision tree using top-down Greedy strategy. This algorithm provides the possibility to create a decision tree based on a fixed set of examples, in order to classify future samples. The tree outputted by this algorithm represents a simple abstraction to explain all the elements of the set and offers in a clever and intuitive way the overall dependencies among them to better understand the system and prepare a decision.

The accuracy of ID3 algorithm can be improved using proposed improved version of ID3 which uses association function along with information gain to decide splitting attribute. This approach also overcomes the shortcoming of choosing multi-valued attributes of ID3 algorithm.

II. ID3 ALGORITHM

The ID3 algorithm is a recursive procedure, which in each step there is an evaluation of a subset and there is the creation of decision node, based on metric called Information Gain, until the subset in evaluation if specified by the same combination of attributes and its values. ID3 algorithm is used to create a decision tree from given set, by using top-down greedy search to check each attribute at every tree node & information gain is used as metric to generate tree.Id3 algorithm uses the information gain as a metrics to select the best attribute in each step.

The ID3 decision makes use of two concepts when creating a tree from top-down:
1. Entropy
2. Information Gain

Using these two metrics, the nodes to be created and the attributes to split on can be determined.
Entropy is a measure in information theory to measure the impurity of a arbitrarily collection of items. For a given set S, being pi the probability of S belonging to class i, we have

$$\text{Entropy H (P)} = -\sum_{(P_i)\ i=1}^{n} P_i \log_2$$

As for Information Gain, we can say that is the expected reduction in entropy by splitting the collection S by a given outcome for attribute A, with an associated subset S_v.

Information Gain uses the entropy in order to determine what attribute is best used to create a split with. By calculating Gain, we are determining the improved entropy by using that attribute. So, the column with the higher Gain will be used as the node of the decision tree.

Information Gain = I (S_1,S_2,S_3,…, S_m)- E(A) Algorithm for generating a decision tree according to a given data sets
Input: training samples, each attribute taking discrete value, a candidate attribute set available for induction is attribute_list.

Output: a decision tree.

1. Create a node *N*.
2. If all samples of the node are of the same category *C*, then return *N* as a leaf node and mark with category *C*, the beginning root node corresponds to all the training samples.
3. If attribute_list is empty, then return N as a leaf node and mark the node as a type whose samples contain the largest number of categories;
4. select a test_attribute with the largest information gain from arrtibute_list, and mark node *N* with test_attribute;
5. For each given value a_i of test_attribute, the sample set contained in node N is portioned.
6. According to the condition of test_attribute = a_i, a corresponding branch is generated from the node N to indicate the test conditions.
7. Set *i s* is the obtained sample set under the condition of test_attribute = a_i. If *i s* is empty, then mark the corresponding leaf node with category of including the most number of sample types. Otherwise, It will be marked with a return value:

(*Generate decision tree* $s_{i,}$ *attribute list − test attribute)*

Experimental Results

A simple loan application dataset is shown in below Table. The category attribute of the sample set is "Class", which will predict whether the new customer's loan application should be approved or not.

CASE	AGE	Has_job	Own_house	Credit rating	Class
1	Young	False	False	Good	No
2	Young	True	False	Good	Yes
3	Young	True	True	Fair	**Yes**
4	Young	False	False	Fair	**No**
5	Middle	False	False	Good	No
6	Middle	True	True	Good	Yes
7	Middle	False	True	Excellent	**Yes**
8	Low	False	True	Excellent	**Yes**
9	Low	False	True	Good	Yes
10	Low	True	False	Good	Yes

Table 1: Training Sample

Both improved ID3 algorithm and ID3 algorithm are applied on this dataset to construct decision trees and comparison is made. Figure 1 and figure 2 show the generated decision trees using the ID3 algorithm and the improved ID3 algorithm, respectively.

Generated Decision tree using given training dataset

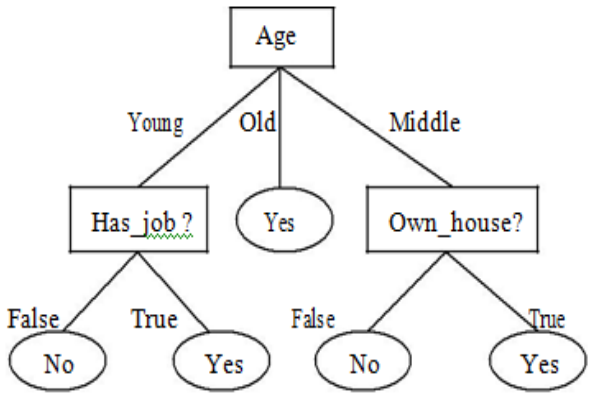

Fig 1: ID3 Decision Tree

Shortcomings of ID3 algorithm

There exist one problem with this approach; this means that id3 selects the attribute having more no. of values, which are not necessarily the best attribute

Data may be over-fitted or over-classified, if a small sample is tested.
Only one attribute at a time is tested for making a decision.
Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

III. IMPROVED ID3

As we have seen and come across the drawbacks of basic ID3 algorithm, so to overcome these shortcomings improved version of ID3 comes into picture. The uneven nature of basic ID3 of selecting multi-valued attributes as a decision node is eliminated by improved ID3.

In Improved version we use the same gain initially calculated in basic ID3 which get changed every time when the dataset get modified as tree grows. In this improved version we first calculate Association function (AF) for each attribute and these obtained values are further used to calculate the normalised gain for each attributes. Now this normalised gain is combined with old (initial) gain of attributes to get new gain for each attribute which is used as standard for making decisions. The schematic methods used for computation of attribute importance are Information entropy, Association function, Normalised gain function etc.

Association function not only very well handles the inadequacy of basic ID3 but also clearly represent relation among elements and attributes.

Computation of AF:

Expermential Results

Figure below shows decision tree generated by using association function:

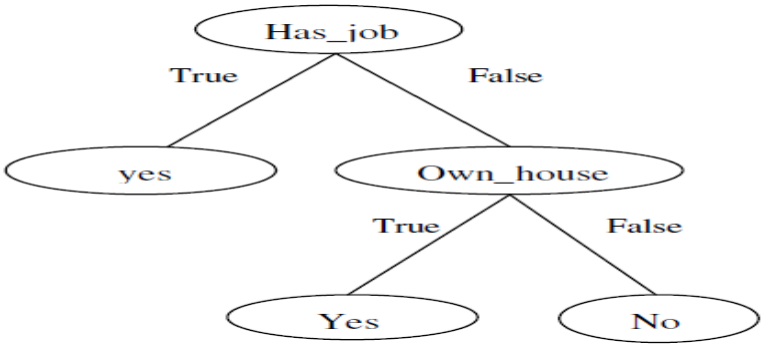

Fig 2: Decision tree using Improved ID3

Suppose attribute belongs to D dataset, and let C be any category attribute of dataset D. Then the relation degree function between A and C can be given as –

$$AF(A) = \frac{\sum_{i=1}^{n} |x_{i1} - x_{i2}|}{n}$$

By closely observing both the trees we found that basic ID3 select Age as root node for making decisions which seems to be worthless and has lower importance, because by selecting age as root for customer loan dataset does not give that much information, but on the other hand decision tree generated by Improved ID3 leads us to efficient and fruitful decision making some sense for given dataset.

Where, x_{i1} is the number of negative attributes X_{i2} is the number of positive attributes

N is number of types of attributes that A contains Now, after calculating AF we calculate Normalised gain for each attribute as follows:

$$V(k) = \frac{AF(k)}{AF(1) + AF(2) + \cdots + AF(m)}$$

Where 0 < K <= m. Further we calculate a new gain for each attribute that allows us to select.
Decision node as

$$Gain'(A) = (I(s_1, s_2, \cdots, s_m) - E(A)) \times V(A)$$

So by following all the above mentioned steps systematically one can overcome the drawbacks of basic ID3.

That means if we consider has job as root then one can get clear cut understanding about weather one can take loan or not. So improved version of ID3 is better than the Basic version.

IV. CONCLUSIONS

1. Accuracy of ID3 algorithm can be improved using association function and more optimal decision trees can be generated using proposed improved ID3 algorithm.
2. In Improved Id3 more reasonable and effective rules are generated.
3. Time complexity is more in improved ID3, but it can be neglected because now faster and faster computers are present.

REFERENCES

[1]. Y. T. Zhang, L. Gong, Principles of Data Mining and Technology, Publishing House of Electronics Industry.
[2]. D. Jiang, Information Theory and Coding [M]: Science and Technology of China University Press, 2001.
[3]. S. F. Chen, Z. Q. Chen, Artificial intelligence in knowledge engineering [M] Nanjing: Nanjing University Press, 1997.
[4]. Z. Z. Shi, Senior Artificial Intelligence [M]. Beijing: Science Press, 1998.
[5]. M. Zhu, Data Mining [M]. Hefei: China University of Science and Technology Press, 2002.67-72.
[6]. J. Quinlan, "Learning decision tree classifiers". ACM Computing Surveys (CSUR), 28(1):71–72, 1996.
[7]. WEKA University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/ (accessed July 18 2012)
[8]. Jiawei Han and Micheline Kamber, "*Data Mining: Concepts and Techniques*", 2nd ed., Morgan Kaufmann Publishers, 2006
[9]. J.R.Quinlan, "*C4.5: Programs for Machine Learning*", Morgan Kaufmann Publishers, Inc, 1992.
[10]. P. Clark and T. Niblett, The CN2 induction algorithm, *Machine Learning* **3** (1989)
[11]. C. C. Chan and J. W. Grzymala-Busse, on the attribute *Report* redundancy and the learning programs ID3, PRISM, and LEM2, *TR-91–14, Department of Computer Science, University of Kansas,* 1991
[12]. I. Kononenko, ID3, sequential Bayes, naive Bayes and Bayesian neural networks, *Proc. of the 4thEuropean Working Session on Learning* (1989) 91–98.
[13]. I. Kononenko and I. Bratko, Information-based evaluation criterion for classifiers performance, *Machine Learning* **6** (1991)
[14]. J. Mingers, An empirical comparison of pruning methods for decision tree induction, *Machine Learning* **4** (1989)
[15]. J. R. Quinlan, Generating production rules from decision trees, *Proc. of the 10th Int. Joint Conf. on AI*, 1987, 304–307