Enhanced ID3 algorithm based on the weightage of the Attribute

[1.] **X.Jose Suganya**.
Research Scholar, JJ College of Arts and Science, Bharathidasan University, Trichy.
Asst.Professor, Shri S.S Shasun Jain College for Women, Chennai-17.

[2.] **Dr. R. Balasubramanian**
Research Guide, JJ College of Arts and Science, Bharathidasan University, Trichy

ABSTRACT - ***ID3 algorithm a decision tree classification algorithm is very popular due to its speed and simplicity in construction but it has its own snags while classifying the ID3 algorithm and tends to choose the attributes with large values and practical complexities arises due to this. To solve this problem the proposed algorithm empowers and uses the importance of the attributes and classifies accordingly to produce effective rules. The proposed algorithm uses the attribute weightage and calculates the information gain for the few values attributes and performs quite better when compared to classical ID3 algorithm. The proposed algorithm is applied on a real time data (i.e.) selection process of employees in a firm for appraisal based on few important attributes and executed.***

Keywords: ID3, Enhanced ID3, Attribute weightage

INTRODUCTION

Perched on anurbane technologicalrealm, the business house’s strategies and promotions have changed dramatically in the recent years since the importance of data in their business plays a pivotal role in almost all of their business activities. Today computer has become a part and parcel of human life and the colossal volume of data available across the globe provides a helping hand to discover useful meaningful information hidden underneath to enhance the business in various activities related to decision making and decision support. Hence the need for an efficient data mining tool or model is imperative for every business organization.Data mining is to discover the relationship and rules hidden in the data, and unearth this precious knowledge hidden in the data and utilize the results to enhance the business.

POSSIBILITIES WITH DATAMINING

Data mining is used in classification, prediction, estimation, association rules, clustering and visualization [1]. Knowledge Discovery in Databases (KDD) is an imperative process of identifying valid interesting, previously unidentifiedbut potentially valuable patterns in data. These patterns enhances the business firms to make predictions or classifications about new data, support decision making and provide graphical data visualization to assist humans in unearthing deeper patterns available.

ID3 ALGORITHM

Quinlan in 1986proposedID3 algorithm [2] is a decision tree learning algorithm based on information entropy and developed from its predecessor named CLS algorithm. Quinlan introduced Shannon's information theory into the decision tree algorithm for calculating the entropy value. The fundamental of ID3 algorithm relies on selecting attributes using information gain as anattribute selection criteria, selecting an attribute with the largest information gain to make decision tree nodes and establishing branches by the different values of the node building the decision tree nodes and branches recursively according to the instances of various branches until a certain subset of the instances belongs to the same category.

For a given a set of examples *S*, each of which is descried by number of attributes along with the class attribute C, the basic pseudo code for the ID3 algorithm is:

```
If (all examples in S belong to class C) then
        Make leaf labeled C
Else
        Select the “most informative” attribute A
        Partition S according to A’s values ($v_1$…..$v_n$)
        Recursively construct sub-trees $T_1$,$T_2$,$T_3$.. $T_n$ for each subset of S.
```

ID3 uses a statistical property, called information gain measure, to select among the candidates attributes at each step while growing the tree. To define the concept of information gain measure, it uses a measure commonly used in information theory, called entropy.

THE ENTROPY IS CALCULATED BY THE FOLLOWING FORMULA,

$$\text{Entropy (S)} = \sum_{i=1}^{c} -pi\ log2\ pi$$

Where *S* is a set, consisting of s data samples, *Pi* is the portion of *S* belonging to the class i. Notice that the entropy is *0* when all members of S belong to the same class and the entropy is 1 when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

In all calculations involving entropy, the outcome of *(0 log₂ 0)* is defined to be 0. With the Information gain measure, given entropy as a measure of the impurity in a collection of training examples, a measure of effectiveness of an attribute in classifying the training data can be defined. This measure is called information gain and is the expected reduction in entropy caused by partitioning the examples according to this attribute.

More precisely, the information gain is calculated by

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{v\ \hat{I}\ values(A)} \frac{|Sv|}{Sv}\ Entropy\ (Sv)$$

Where values of *A* is the set of all possible values for attribute *A*, and Sv is the subset of *S* for which attribute A has value v. The first term in the equation is the entropy of the original collection S, and the second term is the expected value of the entropy after S is partitioned, using attribute *A*. *Gain(S, A)* is the expected reduction in entropy caused by knowing the value of attribute *A*.

Therefore the attribute having the highest information gain is to be preferred in favor of the others. Information gain is precisely the measure used by ID3 to select the best attribute at each step in growing the decision tree.

ENHANCED ID3 ALGORITHM

The main problem in classical ID3 is its nature in choosing attributes with more values and this leads to an incorrect classification on many circumstances. To overcome this snag, this paper introduces an enhanced ID3 algorithm based on attribute weightage and importance. The proposed algorithm utilizes attributes with less values and higher importance whereas alleviates attributes with many values and lesser importance. The experimental analysis clearly showcased that the proposed algorithm produces practical and effective classification rules when compared with ID3 algorithm. Here in the proposed method the importance of the attribute is not determined by its sizebut the selection of the attribute is based on the actual scenario of the dataset and it purely depends on the experienced users who execute the algorithm.

$$0 \leq \rho = \rho\ (A) \leq \text{minimum}\ (\tau_1, \tau_2, \ldots\ldots \tau_n)$$

Where ρ= attribute importance and τ is the probability of attribute A belonging to class C.

Plethora of criteria is to be considered while choosing attribute with lesser values and of higher importance. Considering importance of the attribute purely rely on the current dataset being used and the current scenario, this selection will be carried out by the experienced miners and by the users with exceptional knowledge about the situation. But most of the people are less aware of this fact and ignore the attributes with lesser values and of higher importance due to their inability to access the situation or due to their inferior experience.

The ID3 algorithm formula is modified according to the attribute importance as shown below

$$\text{AIE (A)} = \sum_{i=1}^{N} \frac{((S1i + S2i + \cdots + Sni)}{S} + \mathrm{r}\ (\mathrm{A}))\ I\ (S1i + S2i + \cdots + Sni)$$

Where AIE = Attribute important Entropy

Naturally Gain becomes,

$$\text{Gain(AIE)} = I(S1i + S2i + \cdots + Sni) - AIE(A)$$

Here the user can choose attribute with highest Gain (AIE) as splitting attribute and hence the proposed algorithm employs a new technique in selecting attribute with less values and of higher importance and alleviates the attributes with higher values and of less importance considering the situation.

EXPERIMENTAL EVALUATION

To test the effectiveness of the proposed enhanced ID3 algorithm, a real life dataset corresponding to a software firm to evaluate the performance of the employees for appraisal is considered with 4 attributes. The numerical values provided in the dataset are first converted into general labels to ease the manipulations and calculations.

S.NO	TECHNICAL SKILLS	COMMUNICATION SKILLS	PRODUCTIVITY	QUALIFICATION	APPRAISAL
1	6.8	7.9	8.0	MCA	NO
2	6.1	7.3	7.9	BSc	NO
3	6.5	7.9	7.5	B.E	NO
4	8.1	7.5	7.3	MCA	YES
5	7.3	7.1	7.8	B.E	YES
6	4.2	5.1	7.1	MCA	NO
7	4.7	4.6	7.1	B.E	NO
8	4.3	8.1	6.8	MCA	YES
9	4.4	7.9	6.3	BSc	NO
10	6.7	5.6	7.2	MCA	NO
11	6.4	5.8	6.9	B.E	NO
12	4.5	4.9	6.8	MCA	NO
13	3.9	5.5	6.1	B.E	NO
14	6.3	5.9	6.6	B.E	YES
15	6.7	5.7	6.9	BSc	YES
16	7.3	5.3	7.4	BSc	YES
17	7.8	5.8	7.8	B.E	YES
18	8.1	8.1	6.8	MCA	YES
19	4.2	5.8	7.1	BSc	NO
20	7.7	7.9	6.9	B.E	YES

Table 1: Appraisal dataset with four attributes

The appraisal dataset shown in the table 1 comprises of four attributes namely technical skills of the employee, communication skills of the employee, work productivity of the employee and their qualifications. Note the minimum value and maximum value of skill and productivity are 1 and 10. These values are first labeled or clustered according to the values as shown below.

Technical Skills

Technical skills ≥3< 5 = Average
Technical skills ≥5 < 7 = Good
Technical skills ≥ 7 = Excellent

Communication Skills

Communication skills ≥3 < 7 = Moderate
Communication skills ≥ 7 = Fluent

Productivity

Work Productivity ≥3 < 7 = Normal
Work Productivity ≥ 7 = Great

The Qualification column is already labeled and there is no need to label it further. The numerical scale values present in the appraisal dataset is converted into these labelsand the converted dataset is shown in the table 2.

S.NO	TECHNICAL SKILLS	COMMUNICATION SKILLS	PRODUCTIVITY	QUALIFICATION	APPRAISAL
1	GOOD	FLUENT	GREAT	MCA	NO
2	GOOD	FLUENT	GREAT	BSc	NO
3	GOOD	FLUENT	GREAT	B.E	NO
4	EXCELLENT	FLUENT	GREAT	MCA	YES
5	EXCELLENT	FLUENT	GREAT	B.E	YES
6	AVERAGE	MODERATE	GREAT	MCA	NO
7	AVERAGE	MODERATE	GREAT	B.E	NO
8	AVERAGE	FLUENT	NORMAL	MCA	YES

9	AVERAGE	FLUENT	NORMAL	BSc	NO
10	GOOD	MODERATE	GREAT	MCA	NO
11	GOOD	MODERATE	NORMAL	B.E	NO
12	AVERAGE	MODERATE	NORMAL	MCA	NO
13	AVERAGE	MODERATE	NORMAL	B.E	NO
14	GOOD	MODERATE	NORMAL	B.E	YES
15	GOOD	MODERATE	NORMAL	BSc	YES
16	EXCELLENT	MODERATE	GREAT	BSc	YES
17	EXCELLENT	MODERATE	GREAT	B.E	YES
18	EXCELLENT	FLUENT	NORMAL	MCA	YES
19	AVERAGE	MODERATE	GREAT	BSc	NO
20	EXCELLENT	FLUENT	NORMAL	B.E	YES

Table 2: Normalized Appraisal dataset

The four attributes shown in the table 2 is classified into two classes namely Yes and No. When classical ID3 is executed on this dataset the splitting attribute selected is technical skills since it has more values than other attributes and the decision tree is constructed as shown in the figure 1. This attribute selection is the major disadvantage as far as the classical ID3 is concerned. The classification rules generated from the root node to the leaf nodes will not be a satisfactory one, since in real life scenario there is a need for different and effective rules to provide appraisal to the employees.

DECISION TREE BY ID3 ALGORITHM

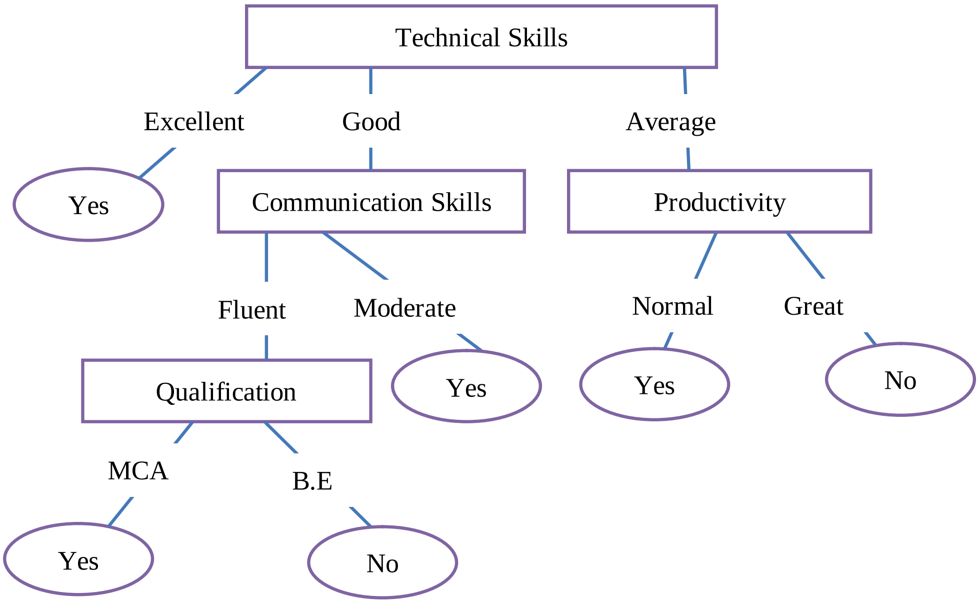

Figure 1: Decision Tree created by ID3 Algorithm

Rule ID	Classification Rules
1	If (Technical skills = Good and Communication Skills = Moderate) THEN (Class = YES)
2	If (Technical skills = Average and Communication Skills = Great) THEN (Class = NO)
3	If (Technical skills = Average and Productivity = Normal) THEN (Class = YES)
4	If (Technical skills = Excellent) THEN (Class = YES)
5	If (Technical skills = Good and Communication Skills = Fluent and Qualification = MCA) THEN (Class = YES)
6	If (Technical skills = Good and Communication Skills = Fluent and Qualification = B.E) THEN (Class = NO)

Table 3: Classification rules generated by ID3

On many occasions the policy makers in the firm believes that the technical skills is more important than any other attribute for appraisal, since it is prestigious for many MNC to employ highly qualified employees rather than scrutinizing the work productivity. But here rule no 4 implies that the technical skill alone will fetch appraisal for the employees. Also from the rule no 3, if the employee possesses average technical skills and normal productivity will fetch appraisal, but the real situation doesn't provide such weird rules.

Here thedecision tree is reconstructed by the enhanced ID3 by diluting the attributes with many values. In that case the technical skills and qualification attributes has threevalues and number of attributes in the dataset is 4.

$$\rho(\text{Technical skills}) = 1 - \frac{\text{Number of values}}{\text{Total attributes}}$$

$$= 1 - (3/4) = 0.25$$

ρ (Qualification) = 1 –(3/4) = 0.25
ρ (Technical skills) = ρ(Qualification) =0.25

The productivity and communication skills attribute has two values
ρ (Productivity) = 1 – (2/4) = 0.5
ρ (Communication skills) = 1 – (2/4) = 0.5
ρ (Technical skills) = ρ(Qualification) = 0.25- 0.25 = 0
ρ (Productivity) = ρ(Communication skills) = 0.5- 0.25 = 0.25

Usual calculations are done with the modified formula, and from the labeled dataset as shown in table 2, there are 9 "Yes" and 11 "No".

The entropy value based on this,

$$I = -\frac{9}{20}\log_2\frac{9}{20} - \frac{11}{20}\log_2\frac{11}{20} = 0.99277$$

INFORMATION ENTROPY

Technical Skills
AIE(TS)= (7/20) (-5/7 $\log_2$ 5/7- 2/7 $\log_2$ 2/7) + (7/20)(-6/7 $\log_2$6/7-1/7 $\log_2$ 1/7) = 0.50917

Productivity
AIE(PR) = (12/20)-0.25(-8/12 $\log_2$ 8/12- 4/12 $\log_2$ 4/12) + (12/20)-0.25(-3/8 $\log_2$3/8-5/8 $\log_2$ 5/8) = 0.46456

Communication Skills
AIE(CS) = (9/20)-0.25(-4/9 $\log_2$ 4/9- 5/9 $\log_2$ 5/9) + (9/20)-0.25(-7/11 $\log_2$7/11-4/11 $\log_2$ 4/11) = 0.48193

Qualification
AIE(Q) = (7/20) (-4/7 $\log_2$ 4/7- 3/7 $\log_2$ 3/7) + (7/20)(-3/5 $\log_2$3/5-2/5 $\log_2$ 2/5) = 0.98756

CALCULATED INFORMATION GAIN

Gain(TS)= 0.99277 – 0.50917 = 0.48360
Gain(PR) = 0.99277 – 0.46456 = 0.52823
Gain(CS) = 0.99277 – 0.48193 = 0.51084
Gain(Q) = 0.99277 – 0.98756 = 0.00521

Here the information Gain of the productivity is found to be the biggest and this attribute selected first for classification and this attribute will naturally become the root of the decision tree. The entire decision tree structure will be different from the classical ID3 decision tree as shown in the figure 2.

DECISION TREE CONSTRUCTED BY ENHANCED ID3

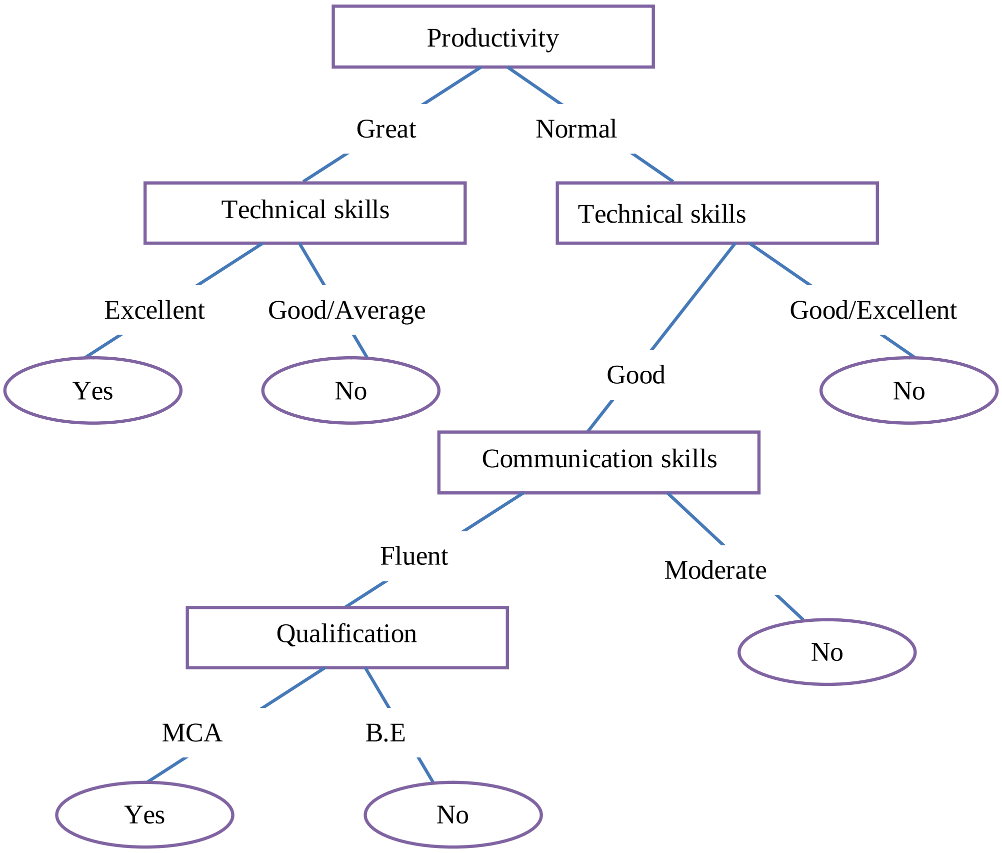

Figure 2: Decision Tree constructed by enhanced ID3

Comparing the two figures 1 and 2, the entire structure of the decision tree is changed and the classical ID3 considered technical skills as the most important attribute whereas the enhanced ID3 considered productivity as the most important attribute. Considering the appraisal in real world situation work productivity plays a pivotal role and the enhanced ID3 has clearly made a better decision than the classical ID3. The rules formed from the enhanced ID3 are shown in table 4.

RULE ID	CLASSIFICATION RULES
1	If (Productivity = Great and Technical Skills = Average / Good) THEN (Class = No)
2	If (Productivity = Normal and Technical Skills = Good and Qualification = MCA) THEN (Class = YES)
3	If (Productivity = Normal and Technical Skills = Average / Good) THEN (Class = No)
4	If (Productivity = Great and Technical skills = Excellent) THEN (Class = YES)
5	If (Productivity = Normal and Technical Skills = Good and Communication skill=Moderate) THEN (Class = No)
6	If (Productivity = Normal and Technical skills = Good and Communication Skills = Fluent and Qualification = B.E) THEN (Class = NO)

Table 4: Classification rules generated by Enhanced ID3

Comparing the classification rules of the two algorithms, it is clear that the enhanced ID3 has given much importance to productivity and technical skill is considered secondary to provide appraisal for employees. The experimental results clearly showcased that the enhanced ID3 has better classification accuracy than classical ID3 algorithm in real world scenario.

CONCLUSION

The enhanced ID3 algorithm employs attribute weightage to discover information entropy to find the information gain of each attribute and splitting occurs with respect to the maximum information gain of the attribute. The enhanced ID3 decision tree and the classification rules matches with that of the real world situation (i.e.) policy makers in the firm gives higher importance to the work productivity followed by the technical skills or expertise, communication skills and finally the qualification. The proposed ID3 algorithm produced results which comply with that of the policy makers.

REFERENCES

[1].Micheal Berry and Gordan ,"Data mining techniques for marketing and customer support", John Willey & sons, 1997.
[2].J. R. QUINLAN. Induction of Decision Trees [J]. Machine Learning.1986,1(1):81-106.
[3].Wei Peng, Juhua Chen and Haiping Zhou. An Implementation Of ID3-Decision Tree Learning Algorithm, School of Computer Science & Engineering, University of New South Wales, Sydney, Australia, http://web.arch.usyd.edu.au/~wpeng/DecisionTre e2.pdf, [Web Accessed 10th july 2006].
[4].Building Classification Models: ID3 And 4.5, http://www.cis.temple.edu/~ingargio/cis58 7/readings/id3-c45.html, [Web Accessed 26th June 2007].
[5].Pardeep Kumar, Nitin, Vivek Kumar Sehgal and Durg Singh Chauhan, ''A BENCHMARK TO SELECT DATA MINING BASED CLASSIFICATION ALGORITHMS FOR BUSINESS INTELLIGENCE AND DECISION SUPPORT SYSTEMS'', International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.2, No.5, pp. 25-42, 2012.
[6].Linna Li, Xuemin Zhang "Study of Data Mining Algorithm Based on Decision Tree", ICCDA IEEE 2010, Pp 78-88.
[7].Chen Jin, Luo De lin, Mu Fen xiang "An Improved ID3 Decision Tree Algorithm" ,IEEE 2009, Pp 76-87.