Hassan A. Jeiad
Computer Eng. Dept
University of Technology,
Baghdad, Iraq.
hsn_uot@yahoo.com

Zinah J. M. Ameen
Computer Eng. Dept.,
University of Technology,
Baghdad, Iraq

Alza A. Mahmood
Computer Eng. Dept.,
University of Technology,
Baghdad, Iraq.

Received on: 31/5/2017
Accepted on: 20/12/2017

Employee Performance Assessment Using Modified Decision Tree

Abstract- *Decision tree algorithms are famous method in inductive learning and successfully applied for model classification and prediction. Performance evaluation in organization is one of the most important issues that are reliable due to the transition from industrial to knowledge age. This paper proposes the use of modified ID3 (Interactive Dichotomiser 3) decision tree algorithm combining with Taneja entropy instead of the original ID3 algorithm that depends on Shannon entropy which is widely used in the information theory. In fact, the original ID3 was suffer from complexity in the form of complex tree with large number of hops and nodes. The information gain was used as a splitting criteria of the modified ID3. The proposed modified ID3 algorithm has been tested on a dataset for a different university employees with several attributes that directly affect their annual performance assessment. The most optimized tree is constructed by taking one attribute that have the largest information gain from the dataset as a root of tree and repeating the process until the tree is completed. The results showed that the proposed modified ID3 decision tree algorithm that based on Taneja entropy gives less complexity due to small tree with three nodes and two to one hope to reach the right decision.*

Keywords- *Data Mining, ID3 algorithm, Entropy, Decision Tree.*

How to cite this article: H.A. Jeiad., Z.J. Mohammed Ameen, and A.A. Mahmood, "Employee Performance Assessment Using Modified Decision Tree," Engineering and Technology Journal, Vol. 36, No.7, pp. 806-811, 2018.

1. Introduction

In today's world, the organizations face many challenges; at the top is the competition among them. To deal with performance assessment in different situations every organization has to take proper and efficient decisions. Most organizations spend the valuable time to steady the plans that make the process of select the proper employees more successful. Commonly, the performance of the leased employees make the managers more concerned, therefore different assessment systems were suggested as a proper approach to seek for the more qualified employees. Data mining that is enables the extraction of unobserved predictive information from enormous databases, is a promised technology with major potential that help in condensing on most paramount information [1, 2]. There are different techniques to mine the data from databases, such as association rules mining, clustering and classification and segmentation, under which decision trees are formed. Mainly, the classification techniques can be considered as a supervised learning techniques introduced to classify data item into different class labels that were predefined. Decision tree is one of the predominant techniques widely used in the data mining because it makes the decision from the data given utilizing clear equations depending principally on computation of gain ratio, which uses some sort of metrical attributes, and then the attribute with a highest gain value would be the more valuable attribute on the anticipated target. As a result, a certain decision tree would be suggested and created with a well-defined classification rule [2]. One of the most famous decision trees is ID3 (Interactive Dichotomiser 3) decision tree algorithm that is widely used in the performance assessments and decision making topics within the data mining field. Essentially, the core calculation of ID3 algorithm is based on Shannon entropy formula that is very well-known formula in the information theory. In fact, the standard ID3 was suffer from complexity in the form of complex tree with large number of hops and nodes [3, 4]. This paper proposes the use of modified ID3 decision tree algorithm that based on Taneja entropy formula instead of the original ID3 algorithm that depends on Shannon entropy. Commonly, the formula of Taneja entropy is belong to α and β classentropy and contains more information than Shannon's entropy and α-class entropies such as Havrda and Charvat entropy [9].

https://doi.org/10.30684/etj.36.7A.14
2412-0758/University of Technology-Iraq, Baghdad, Iraq

This paper is organized as follows: Section (2) presents the related work. Section (3) describes the standard ID3 algorithm in general. Section (4) presents the proposed modified ID3 decision tree. Section (5) introduces the implementation and experimental result with numerical examples. Finally, the results analyzed in section (6), while some conclusions were presented section (7).

2. Related Work

Hssina et al. [3] introduced a comparative study of decision tree ID3 and C4.5 then made a differentiation between those two algorithms and others such as C5.0 and Bhardwaj [4] implemented ID3 decision tree learning algorithm on weather datasets for playing crickets. Hazra et al. [5] presented a model that helps organizations to choose the candidates efficiently and within a short period based on relevant attributes and using ID3 for decision tree building. Mathur et al. [6] used Havrda and Charvat entropy instead of Shannon Entropy to take the better decision in analyzing the data. Jabbar et al. [7] proposed a system based on texture features classification for multi object images by using ID3 decision tree algorithm. The proposal uses image segment tile base to reduce the block effect and uses Wavelet Haar to reduce image size without loss of any important information. Abdulmunim et al. [8] designed a system to read a mammogram image to diagnose the breast cancer by using wavelet and contourlet transform based on various operations on mammogram images and classifies them depending on the decision tree ID3 algorithm.

In this paper Taneja formula was proposed to be a suitable alternative for Shannon formula to introduce a modified ID3 algorithm that simplifies the decision tree with less number of nodes.

3. Standard ID3

The idea of decision trees was introduced and refined over numerous years by J. Ross Quinlan beginning with ID3 (Interactive Dichotomizer 3) [9]. The ID3 algorithm starts with the first set S as the root node. On every each restoration of the considered algorithm, it is repeated through each unused attribute of the set S and figures the entropy E(S) and information gain of that attribute. It then chooses the attribute which has the littlest entropy (or greatest information gain) value. The set S then parts by the chose attribute to deliver subsets of the data. The algorithm repeats itself on every subset, considering the attributes never been chosen before [10]. Figure 1 explains the ID3 algorithm.

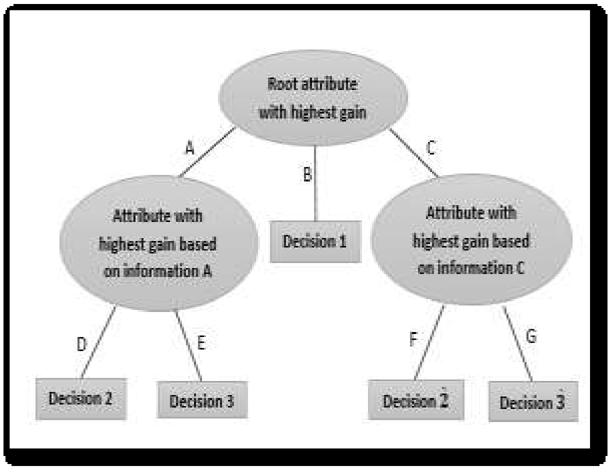
Root attribute with highest gain
A
B
C
Attribute with highest gain based on information A
Decision 1
Attribute with highest gain based on information C
D
E
F
G
Decision 2
Decision 3
Decision 2
Decision 3

Figure 1: ID3 Algorithm

Entropy E(S) is a basic concept in physics and information science, being the basic measure to compare different states of an isolated system (the information content of a description). Entropy measures the degree of our lack of information about a system [10]. Entropy utilizes information gain as its attributes choice measure. The attribute with most noteworthy information gain is picked as the splitting attribute for a given node. The aim is to search for the feature that best splits the current class into child nodes according to their target classification [11].

4. Proposed Modified ID3

In this section, the proposed modified ID3 decision tree algorithm have been introduced through determining both of the entropy and the information gain depending on Taneja entropy formula, and then building the corresponding decision tree. Firstly, Shannon entropy formula and Taneja entropy formula were depicted and when we reached how to build the decision tree both of these formulas were calculated for comparison purposes. After that, the information gain calculation was illustrated to be used in the building of the decision tree. So, both of the entropy formulas were depicted as follows:

I. Shannon Entropy

Shannon defines entropy in terms of a discrete random variable of set S, with possible states s1...sn as:

$$E(S) = -\sum_{i=1}^{n} p(si)\log_2 p(si) \qquad (1)$$

Where $p(si)$ is the probability of S. Shannon entropy is the common entropy formula that is widely applied in many different fields, especially in building the ID3 decision tree, but in some cases, it could lead to a complex tree with many nodes and leafs that makes the decision process delayed and consumed much time especially for big datasets [12]. Due to those limitations, this work proposes the use of Taneja entropy to calculate information gain of the defined attributes instead of Shannon entropy to overcome the main limitation of tree complexity.

II. Taneja entropy

Tis entropy formula which belongs to α and β-class entropy contains more information than Shannon's entropy and α-class entropies such as Havrda and Charvat entropy [12]. The parametric generalized measures of entropy for a given normal distribution are more advantageous as they provide more information about the density function than Shannon's entropy. Taneja entropic form is [13]:

$$E(S) = \frac{1}{1-\alpha}\sum_{i=1}^{n}\frac{si^{\beta+\alpha-1}}{si^{\beta}} \qquad (2)$$

Where α and β are the inherent parameter of the set S.

III. Information Gain

The information gain that represents the amount of information needed after splitting the tree was determined. The information gain of an Attribute Att. is measured using Equation 3:

$$Gain(S, Att.) = E\ (S) - \sum_{v\ \in Values\ (Att.)}\frac{|si|}{|s|}Entropy\ (sv)$$

(3)

Where $Values\ (Att.)$ Denotes to the set of all possible values of attribute *Att*, si are possible states of the set *S*.

5. Implementation and Experimental Results

Building the Decision tree needs to find which attribute would be the root node of the tree. The entropy and the information gain are calculated depending on the proposed training set to resolve the splitting aspect for building the decision tree. Entropy value is computed using Shannon and Taneja entropy formulas as previously mentioned for the comparison purposes. The sample dataset consists of 12 records of different university employees' main attributes that directly affect their annual performance assessment. The dataset comprises of different qualitative and quantitative features. The attributes "ID", "Department", "Activity", "Publication", "Load", "Bonus", and "Other Work" were taken into consideration. The quantitative aspects are "Department" and "Other Work", on the other hand the qualitative skills are "Activity", "Bonus", "Publication" and "Load". Table 1 indicates the training datasets attributes with the possible values. ID and department values have no effect on decision making therefore they are excluded. On the basis of proposed training dataset the entropy and the information gain are computed to resolve the dividing aspect for building the decision tree to determine the employee final performance evaluation that would be either very good (V.G), good (G) or bad (B).

Now, we will go to calculate the entropies for both of Shannon and Taneja entropies and information gain for the purpose of generating the modified ID3 decision tree. The calculation is done in the next two steps as follows:

Step1: Calculation of Shannon entropies and information gain.

The suggested dataset have 12 employees with 5V.G, 3G and 4B degrees, so, E(S) can be calculated using Equation 1 as follows:

$$E(S) = -\frac{5}{12}\log(\frac{5}{12}) - \frac{3}{12}\log(\frac{3}{12}) - \frac{4}{12}\log(\frac{4}{12}) = 0.4659$$

For the first attribute which is called activity that has three possible values 0, 1-2 and 3-5 as illustrated in Table 1. Activity (0) has 2B, 1G and 1 V.G degrees so the entropy for activity (0) is

$$E(0) = -0.5\log(0.5) - 0.25 \times 2\log(0.25) = 0.4515$$

Activity (1-2) has 2B, 2G and 1 V.G. degrees, its entropy is computed as follow:

$$E(1-2) = -\frac{1}{5}\log(\frac{1}{5}) - \frac{2}{5} \times 2\log(\frac{2}{5}) = 0.458$$

Finally activity (3-5) has 3 V.G degrees only, so it would be calculated as: $E(3-5) = 0$

So, activity information gain is computed as follow using Equation 3:

$$Gain(S, Activity) = 0.4659 - \left[\frac{5}{12} \times 0.458 - \frac{4}{12} \times 0.4515\right] = 0.1246$$

By repeating the previous steps for each of the other four attributes, we get the results in Table 2.

Table 1: Description for the used Dataset

Attributes	Possible Values	Description
Department	Information Engineering (IE), Network Engineering (NW)	Branches of the employee
Activity	0, 1-2, 3-5	Participating in conferences or other seminars inside or outside the college
Publication	Yes, No	If the Employee has published
Load	0, 1, 2-3	Teaching Load
Bonus	0, 1-3	Honors or awards received
Other Work	Yes, No	Other departmental work
Rank	Very Good(V.G.), Good(G), Bad(B)	Final assessment Degree

Table 2: Root node calculation using Shannon entropy.

Attribute	Information gain
Gain(S, Activity)	0.1246
Gain(S, Publication)	**0.2701**
Gain(S, Load)	0.0733
Gain(S, Bonus)	0.1023
Gain(S, Other work)	0.057

As indicated in Table 2, publication attribute has the highest gain value. The attribute that have biggest gain value would be the root node according to ID3 algorithm. Publication attribute has only two possible values (yes or no) as shown in Table 1. By splitting the main dataset into two tables according to root attribute given values and repeat the previous calculation we get a new E(S) and information gain for each split. Until we get the final decision tree as shown in Figure 2.

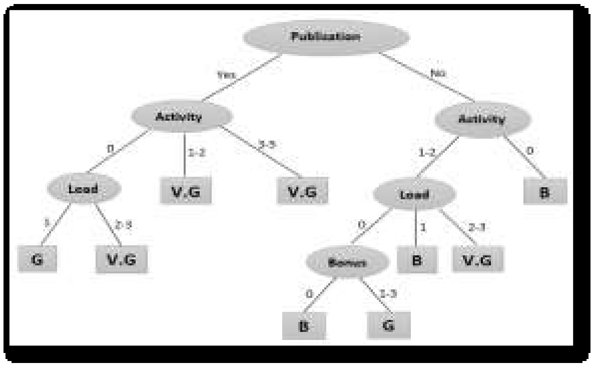

Figure 2: Decision tree using Shannon entropy.

Step 2: Calculation of Taneja entropies and information gain

Initially, we assumed that the values of α and β in Equation 2 are taken to equal 0.75 and 0.5 respectively. Depending on the same scenario applied in step (1), E(s) is calculated as follow:

$$E(S) = \frac{1}{1-0.75} \times \frac{(\frac{5}{12})^{0.25}+(\frac{3}{12})^{0.25}+(\frac{4}{12})^{0.25}}{(\frac{5}{12})^{0.5}+\left(\frac{3}{12}\right)^{0.5}+(\frac{4}{12})^{0.5}} = 5.2712$$

Taking the first attribute (activity) with its possible values:

$$E(0) = \frac{1}{1-0.75} \times \frac{(0.25)^{0.25}+ (0.5)^{0.25}}{(0.25)^{0.5}+ (0.5)^{0.5}} = 5.284$$

$$E(1-2) = \frac{1}{1-0.75} \times \frac{(\frac{1}{5})^{0.25}+2\times(\frac{2}{5})^{0.25}}{(\frac{1}{5})^{0.5}+2\times(\frac{2}{5})^{0.5}} = 5.2785$$

$$E(3-5) = \frac{1}{1-0.75} \times \frac{(\frac{3}{3})^{0.25}}{(\frac{3}{3})^{0.5}} = 4$$

So, activity information gain is computed using Equation 3 as follows:

$Gain(S, Activity) = 5.2712 - \left[\frac{5}{12} \times 5.2785 + \frac{3}{12} \times 4 + \frac{4}{12} \times 5.284\right]$ =0.3105 By repeating the previous steps for each attributes given possible values, we obtained the results in Table 3.

As indicated in Table 3, activity attribute has the highest gain value. So, activity attribute is the root node. It is clear that the activity attribute contains three different ranges which are 0, 1-2 and 3-5 as shown in Table 1. By splitting the main dataset into three tables according to root attribute given values and repeat the previous calculation we get a new E(S) and information gain for each split, till we get the final decision tree as shown in Figure 3.

Table 3: Root node calculation using Taneja entropy.

Attribute	Information gain
Gain(S, Activity)	**0.3105**
Gain(S, Publication)	0.283
Gain(S, Load)	0.1026
Gain(S, Bonus)	0.1887
Gain(S, Other work)	0.075

Other values for α and β parameters were taken into consideration, such that when α=2 and β=1, decision tree will be more complicated and similar to Shannon entropy decision tree whereas using α=0.3 and β=0.6, the decision tree will be simpler with shorter path.

It is obvious that the decision tree in Figure 3, which is built using modified ID3 algorithm based on Taneja entropy is simpler with shorter path from the root node to the leafs than the other decision tree built using ID3 algorithm based on Shannon entropy which is shown in Figure 2.

The four datasets in Table 4 are used to predict the rank (marked by question symbol (?) in the last column) of the employee using the decision trees shown in Figures 2 and 3 for testing purposes. We need to pursue the two trees to predict the ranks For the purpose of filling the last column of Table 4. For employee with ID (13) we will find that this employee will gets bad degree based on first tree in Figure 1 with 4 hops. The same employee will take the same degree with 2 hops only by utilizing the second decision tree in Figure 3. In the same way employee with ID (14) will take very good. Employee with ID (15) will have good degree and employee with ID (16) will have very good degree.

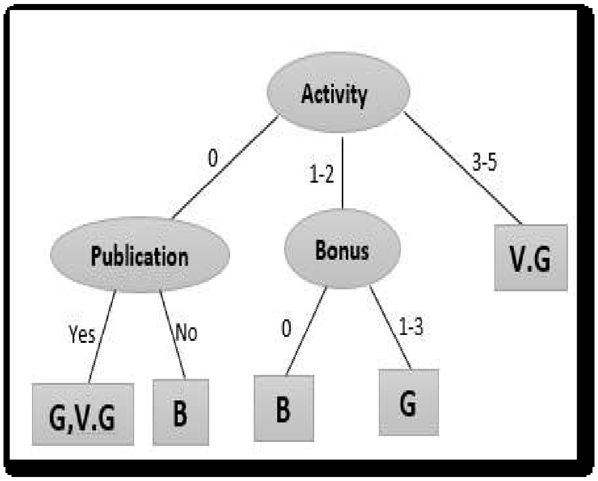

Figure 3: Decision tree using Taneja entropy.

6. Result Analysis

The results that were obtained from the previous section showed that using Taneja entropy formula within the modified ID3 algorithm to find the decision tree can give the ability to calibrate the complexity of that decision tree comparing to the fixed form of tree complexity obtained through using Shannon formula. The calibration ability is came from adjusting the values of the parameters α and β in Taneja formula to produce a decision trees with different complexity levels. The results showed that assigning values of α=0.3 and β=0.6 gives a decision tree with less node and hopes than that for α=0.75 and β=0.5. The number of nodes and hopes will be increased to be equal to that obtained from the standard ID3 algorithm when α=2 and β=1. In general, assigning values for α and β greater or equal to 1 will makes both of the standard and the modified ID3 algorithms to give the same complexity of decision tree, whereas, values less than 1 for α and β will guarantees the modified ID3 algorithm to produce decision tree with less complexity comparing to the standard ID3 algorithm.

7. Conclusions

This paper has concentrated on the ability of proposing a classification model for predicting the evaluation of employees performance depending on a modified ID3 decision tree algorithm that based on Taneja entropy formula. A dataset with different attributes have been examined. The standard ID3 algorithm applied using two entropy formulas which are Shannon entropy Taneja entropy that is called as (α,β)-class entropy. Thus, two different decision trees were established from which we found that some attributes influence the performance assessment prognosis and others might not. The Modification of ID3 algorithm was made using Taneja entropy in the calculation of the information gain and choosing the dividing attributes. It has been observed that by applying values for α and β less than 1 in Taneja entropy, the decision tree will be simpler with less number of nodes. Otherwise, using values for α and β greater than 1 the tree will be closer to the tree established using Shannon entropy.

Thus, adjusting the values of α and β for can adjust the deep and complexity of decision tree. Therefore, applying this model in companies and organizations management for follow up employees and their performance will be so suitable.

Table 4: Testing Dataset

ID	Department	Activity	Publication	Load	Bonus	Other Work	Rank
13	NW	1	No	0	0	Yes	?
14	IE	3	Yes	0	0	Yes	?
15	NW	1	No	3	2	No	?
16	IE	5	Yes	0	0	Yes	?

8. References

[1] H. Chahal, "ID3 Modification and Implementation in Data Mining," International Journal of Computer Applications (0975-8887), Vol. 80-No7, 2013.

[2] O.A. Al-Radaideh, and E. Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," International Journal of Advanced Computer Science and Applications, Vol. 3 No. 2, 2012.

[3] P. Kumar, and D.H. Hooda, "On Generalized Measures of Entropy and Dependences," Mathematica Slovaca Journal, Vol. 58, No. 3, 2008.

[4] G. Maksa, "The Stability of the Entropy of Degree Alpha," Elsevier Journal, Vol. 346, No. 17-21, 2008.

[5] S. Hazra, and S. Sanyal, "Recruitment Prediction Using ID3 Decision Tree," International Journal of Advance Engineering and Research Development, Vol. 3, Issue 10, 2016.

[6] N. Mathur, S. Kumar, and R. Jindal, "The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree," International Journal of Information Engineering, Vol. 2, No. 2, 2012.

[7] E.K. Jabbar, and M.J. Kelain, "Classification of Images Using Decision Tree," Eng. & Tech. Journal, Vol. 31, Part B, No. 6, 2013.

[8] M. E. Abdulmunim, and Z. F. Abed, "Classification Mammogram Images Using ID3 decision tree Algorithm Based on Contourlet Transform," Eng. & Tech. Journal, Vol. 33, Part B, No. 3, 2015.

[9] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A Comparative Study of Decision tree ID3 and C4.5," International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 2013.

[10] P. Rud, "Data Mining Cookbook," John Wiley & Son, Inc, 2001.

[11] J.R. Watson and E.M. Carson, "Undergraduate students' understandings of entropy and Gibbs free energy," University Chemistry Education, Vol. 6, 1, 2002.

[12] P. Eastman, "Introduction to Statistical Mechanics," John Wiley & Son, Inc, 2014.

[13] R. Bhardwaj, and S. Vatta, "Implementation of ID3 Algorithm," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, issue 6, 2013.

[14] N. Magesh, P. Thangaraj, S. Sivagobika, S. Praba, and R. M. Priya, "Employee Performance Evaluation using Machine Learning Algorithm," International Journal of Computer Communication and Networks, Vol. 4, No. 2, 2014.

[15] N. Venkatesan, K. Arunmozhi Arasan, and S. Muthukumaran, "An ID3 Algorithm for Performance of Decision Tree in Predicting Student's Absence in an Academic Year using Categorical Datasets," Indian Journal of Science and Technology, Vol. 8, 14, 2015.

Author's biography

Hassan Awheed Jeiad: Received the B.Sc. degree in 1989 in Electronics and Communications Engineering from University of Technology, Baghdad, Iraq. He is received the M. Sc. degree in Communication Engineering from University of Technology, Baghdad, Iraq in 1999. He is received the Ph.D. in 2006 in Computer Engineering from University of Technology, Baghdad, Iraq. He is currently a lecturer in the Department of Computer Engineering in the University of Technology, Baghdad, Iraq. His research interests include computer architecture, microprocessors, computer networks, multimedia, adaptive systems, and information systems.

Zinah Jaaffar Mohammed Ameen: Received the B.Sc. degree in 2005 in Software Engineering from University of Technology/ Baghdad. Received the M. Sc. degree in Information Engineering from Al Nahrain University/Baghdad in 2011. She is currently an assistant lecturer in Computer Engineering department in University of Technology, Baghdad, Iraq. Her research interests include Computer network, Information Technology, and Data Mining.

.

Alza Abduljabbar Mahmood: Received the B.Sc. degree in 2006 in Computer Engineering from University of Technology/ Baghdad. Received the M. Sc. degree in Information Engineering from Al Nahrain University /Baghdad in 2012. She is currently an assistant lecturer in Computer Engineering department in University of Technology, Baghdad, Iraq. Her research interests include Computer networks, Communication, Information systems and Data Mining.