Modification of ID3 Algorithm to Explore Students' Performance with their Attributes

NS Rathnayaka[1], JK Wijerathna[2]

[1]*Department of IT and Mathematics, General Sir John Kotelawala Defence University, Sri Lanka*
[2]*Department of Mathematics, Faculty of Science, University of Colombo, Sri Lanka*
[1] nsrathnayaka@gmail.com, [2] jagath@sci.cmb.ac.lk

Astract- *Classification is one of the influential techniques in data mining process which can apply to accurately predict the target class for each case in a data set. Particularly, in the field of education the object of classification is the categorization of data for its most effective and efficient utilization. The Decision tree is one of the important methods of classification in data mining for facilitating decision making, in sequential decision problems from historical data. Iterative Dichotomiser 3 or ID3 is the one of the popular decision tree algorithm invented by J. Ross Quinlan in 1986. This algorithm does not make the better decision rules for some data sets. In this paper a new modified decision tree algorithm is proposed to overcome the problem with the conventional ID3 algorithm, so that to choose the attribute with many values. The experimental result shows that, the proposed algorithm performs better than the formal algorithm and is implemented using MATLAB.*

Keywords: *Data mining, Classification, Decision Tree algorithm, Accuracy*

I. INTRODUCTION

The Collecting of huge amount of data is rapidly increasing in many areas, including Customer Relationship Management, Marketing, Medical Diagnosis, Science, Law, Education, web mining in year by year. This historical data set may include some information that can be used to amend the future consequences. Data mining some time called the Knowledge discovery in databases (KDD) is a process used by data minors, to covert this raw data into useful information (Han & Kamber 2006) (M.M.Tom 1997). Nowadays, researchers are interesting for data mining in education. These new imaging field is identified as Educational data mining (EDM), concern with developing the method that discovers the knowledge from data originating from the educational environment. The main objective of any educational institute is to provide quality education to their students and the students' objective is to obtain higher grades in the respective subjects (K.B.Brijesh & S.Pal 2011).

In prospect of data mining, there are several techniques called associations, clustering, classification, modeling, sequential patterns, and time series forecasting. The Decision tree is a one of powerful algorithms for classification and prediction which facilitates to make the decision in sequential decision problems of the any kind of dataset. There are many decision tree classification algorithms such as ID3, C4.5, CART, and CHAID, etc. Merely, it is being identified that, the error classification has become the major issue in term of good prediction accuracy of these decision tree algorithms for some data set. Hence, the objective of this work is to identify the shortcoming of ID3 algorithm using the students' data set and present the associated function to improve the decision rule of ID3 algorithm.

II. RELATED STUDIES

Decision tree algorithms are more popular and powerful classification technique in data mining. Because of that, the data miners are motivated to utilize the decision tree algorithm as their data mining tools.

The researchers, B. Aashoo, D. Kavita and S.Manish (2012), conducted to study under the topic on the implementation of decision trees (B.Aashoo, D.Kavita & S.Manish 2012). This paper introduces some modification over the traditional algorithm ID3 and C4.5 to make capable the algorithms to work with large dataset with higher performance. The paper has been concluded that the designed modification is less time consuming for large data set but this has not completely solved the problem to obtain tree with higher performance.

Another improvement of ID3 algorithm is done by S. Anumitha, S. Diana and D. Suganya in 2012. They conducted studies under the topic on Improvisation of ID3 Algorithm Explored on Wisconsin Breast Cancer Dataset (S.Anumitha, Diana & D.Suganya 2012). In this paper, an improvised version of ID3 was experimenting using Wisconsin Breast Cancer dataset. The Breast Cancer dataset was used for executing five Feature Selection algorithms, namely ReliefF, StepDisc, Forward-Logit, Backward-Logit and Fisher Filtering. Using the features selected dataset; better Decision Trees are obtained by running the improvised ID3 algorithm.

In 2013, V. Maduskar and Y. Kelkar, conducted study A New Modified Decision Tree Algorithm based on ID3 (V.Maduskar & Y.Kelkar 2013). The proposed new modified decision tree algorithm combines the concept of similarity measure and found that the new algorithm is compute better accuracy than conventional algorithm and the memory uses of modified ID3 is higher than ID3.

R. Bhardwaj and S. Vatta (2013), conduct study on implementation of ID3 algorithm (B.Rupali & S.Vatta 2013). In this paper the ID3 decision tree learning algorithm is implemented with the help of an example which includes the training set of two weeks. The studies and their implementation conducted that the decision tree learning algorithm ID3 works well on any classification problems having dataset with the discrete values. In this paper we proposed a new technique to improve the Information gain in ID3 algorithm.

III. CLASSIFICATION AND PREDICTION

Data mining algorithms can follow three different learning approaches as supervised, unsupervised as well as semi-supervised. In the supervised learning method the algorithm works with a set of examples whose labels are known. In these learning methods the attributes are nominal in the case of the classification task and attributes are continuous in the case of the regression task. But in an unsupervised learning the labels of the instances in the dataset are unknown and aims of the algorithm is group the examples according to the similarity of their attribute values, characterizing a clustering task. Semi-supervised learning is usually used a small subset of labeled examples together with a large number of unlabeled examples (S.Beniwal & J.Arora 2012).

Classification is also called supervised learning which can be mainly proceeding in a two step process: model construction and usage of the constructed model. Initially, concentrate to build the classification model by describing a set of predetermine classes from a training set as a result of learning from that dataset. Each sample in the training data set is prefigured to comprehend to a predefined class, as determined by the class attribute label (O.D.Samue 2006). Then classify future or unknown objects based on the patterns observed in the training set and then estimate the accuracy of the built model. Under the several data classification techniques we concentrate only the ID3 decision tree classification algorithm.

IV. DECISION TREE

A decision tree (DT) is a graphical representation in a tree structure of every possible outcome of a decision, which includes three kinds of nodes and branches. The DT classifier is built in two phases, Growth phase and Pruning phase. Initially DT is building by recursively splitting the training set of data. While growing, the tree may be over fit data. Then the pruning phase handles the problem of over fitting of the data in the DT (S.K.Yadav, B.Bharadwaj & S.Pal 2012). The prune phase generalizes the tree by removing the noise and outliers and it will increase the accuracy of the decision tree.

A. ID3 (Iterative Dichotomiser 3)

In the decision tree learning, ID3 is an algorithm developed by computer science researcher J. Ross Quinlan in 1986. He used two basic concepts for determining the relationship among the data called entropy and information gain. The attribute with highest information gain measure is taken as a splitting attribute. ID3 creates understandable prediction rules from the training dataset and also it builds quite the faster and shorter decision tree (S.Anumitha, Diana & D.Suganya 2012).

The main method of the ID3 is based on the Concept Learning System (CLS) algorithm. i.e Let C be the set of training examples,

Step 1: If all examples in C positive then create, True node and stop. Moreover, if all examples in C are negative, create a false node and stop.

Step 2: Divide the training examples in C into subsets $C_1, C_2, ..., C_n$ according to their corresponding values.

Step 3: Recursively do the algorithm for each of the sets C_i.

ID3 explores all attributes and select the perfectly separates attributes of the set of training examples. If the attribute can classify perfectly in the training set, then ID3 will stop. Else if, it recursively partition subsets to find out the best attribute. The greedy approach is used in ID3. That is it will forward for the best attribute and never turn back to reconsider the previous decision.

B. Information Gain

ID3 use the Information gain as its attribute selection measure. This is most popular impurity functions used for decision tree learning algorithm and is based on the entropy function from information theory (Han & Kamber 2006). Entropy measures the amount of information in an attribute.

The expected information (entropy) is the total information needed to classify a tuple/record in D is defined as:

$$Info(D) = -\sum_{i=1}^{m} \frac{|C_{i,D}|}{|D|} \log_2\left(\frac{|C_{i,D}|}{|D|}\right) \quad (1)$$

where $|C_{i,D}|/|D|$ is the probability that an arbitrary tuple in D belongs to class C_i.

In order to classify a tuple from D based on the partitioning by attribute A_i, having distinct values, $\{a_1, a_2, a_3, \ldots, a_v\}$, as observed from the training dataset and attribute A_i will be used to divide the data set D into $|v|$ partition or subset, $\{D_1, D_2, D_3, ..., D_v\}$, where D_i contains those tuples in D that have outcome a_j of A_i, need to find the conditional entropy defined as:

$$Info_{A_i}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), \quad (2)$$

where the term $|D_j|/|D|$ act as the weight of the j^{th} partition. Then the information gain value of the attribute A_i is defend as the difference of above two values.

$$InfoGain\ (D, A_i) = Info(D) - Info_{A_i}(D), \quad (3)$$

$InfoGain\ (D, A_i)$ tells us how much would be gained by branching on A. it is the expected reduction in the information requirement caused by knowing the value of *A*. The attribute A with highest information gain ($InfoGain\ (D, A_i)$), is chosen as a splitting attribute at node *N* (Han & Kamber 2006).

This ID3 algorithm is often biased to select the attributes with more taken values in an attribute, but which is not of necessity the best attributes for decision making. So we can overcome this problem by improving the ID3 algorithm.

V. THE IMPROVED ID3 ALGORITHM

To overcome the above mentioned shortcoming, the associated function (4) values at each attribute and information gain should be combined together. Then create the new standard method for attribute selection to build the decision tree (C.Jin, L.De-Lin & M.Fen-xiang 2009).

Supposed *A* is an attributes of the data set *D*, and *C* is the categorical attribute of *D*, then the relation degree function between *A* and *C* can be expressed

$$AF(A_m) = \frac{\sum_{i=1}^{n} (x_{i1} - x_{i2})^2}{n} \quad (4)$$

Where x_{ij} (*j*= 1, 2 represent two kind of class values) indicates that attribute *A* of *D* takes the i^{th} value and category attribute *C* takes the sample number of the j^{th} value, *n* is the number of values of an attribute *A* takes and *m* is the number of attributes of the dataset *D* (S.Anumitha, Diana & D.Suganya 2012).

Then, the normalization of relation degree function value is:

$$V(A_k) = \frac{AF(A_k)}{\sum_{i=1}^{m} AF(A_i)}, \quad (5)$$

Where $0 < k \le m$ and $k \in N^+$.

If *s* is the number of successes in class related to the attribute A, then, the improved version of information gain, $IInfoGain(A)$ formula is defined as:

$$IInfoGain(D, A_k) = (Entropy(D) - Entropy_{A_k}(D)) \times V(A_k) \times |n - s| \quad (6)$$

This value can be used for new attribute selection criteria and it will overcome the shortcoming of the classical ID3 algorithm.

VI. DESCRIPTION OF THE DATASET

This data set includes the students' information with the academic performance of the Faculty of Engineering at KDU, Sri Lanka. The size of the dataset is 37 tuples with 7 attributes. The data were gathered through a questionnaire. The predicted attribute of the dataset is "students' performance", which considered two different values, SGPA greater than or equal 3.00 and SGPA less than 3.00. The coding systems were introduced for the each numerical as well as nominal attributes of the data set. Then, MATLAB program was written to implement the decision tree algorithm to determine the splitting attributes. The coding system of the data set is represented in the Table 1.

Table 1. Cording of the attributes

Students' Performance (SGPA)		**Z-Score (A/L)**		**Dept.**		**Family Income**		**School type (A/L)**		**self studies hours per day**		**No of Games**		**Parents are in academic field**	
*SGPA >= 3.00	14	x>1.00	4	MR	5	medium	9	National School	12	x<=2	20	G- 1	16	Yes	23
*SGPA< 3.00	15	0.9<x<=1	3	CE	6	low	10	other	13	2<x<=3	21	G-2	17	No	24
		x<=0.9	2	EE	7	high	11			x>3	22	G-3	18		
				ME	8							G-4	19		

*SGPA- Semester grade point value

VII. EXPERIMENT RESULT

The objective of this research is to, identify deficiency of ID3 decision tree algorithm and improve the classical algorithm to overcome the shortcoming. The result is obtained by the MATLAB program for the classical and improved ID3 algorithms (IID3). The obtained results are discussed in the following two sections. The information gain values for the classical ID3 algorithm and the improved information gain values for the IID3 algorithm for each attributes are shown in the Table 2.

A. Classification Result Obtained From Classical ID3 Algorithm

The Figure 1 is obtained by using ID3 algorithm for the preprocessed dataset with 37 tuples and 7 attributes. The height of the decision tree is 3 and is not so complex. The family income attribute is preferred as a root note by the classical ID3 algorithm. And so, the branch indicated by 9 (median income) and 10 (low income) are selected the department attribute as a splitting parent node. But importance of this attribute is lower than to the other attributes to make the decision rule to predict the students' SGPA value.

B. Decision Tree For Improved ID3 Algorithm

The Figure 2 is obtained by using IID3 algorithm for the same data set which used for obtaining the ID3 decision tree.

The height of the decision tree which is obtained by improved ID3 algorithm is 5 and then the tree is complex than the decision tree in Figure 1. Root node of the decision tree is family income and it's same as Figure 1. It splits into three branches by indicating the numbers (medium income), 10 (low income), and 11(high income). Medium income branch split to the two event based on parent s' academic involvement and it further split based on z-score. Hence this decision tree has been selected important attribute to predict the students' SGPA values value and it is the overcome of the shortcoming of the ID3 algorithm.

Table 2. Attributes splitting information

Attributes	AF	V	Conditional Entropy	Info Gain	V*Info Gain	V*Info Gain *|n-s|
(A/L) Z-Score	28.6667	0.1212	0.8591	0.0904	0.0110	0.2300
Department	9.0000	0.0381	0.9306	0.0188	0.0007	0.0143
No of games	25.5000	0.1078	0.8726	0.0768	0.0083	0.1657
Family Income	49.3333	0.2086	0.8213	**0.1281**	0.0267	**0.5613**
School type (A/L)	61.0000	0.2579	0.9167	0.0328	0.0085	0.1861
Self Study hours	22.0000	0.0930	0.9039	0.0455	0.0042	0.0890
Parents are in academic	41.0000	0.1734	0.8690	0.0804	0.0139	0.3067

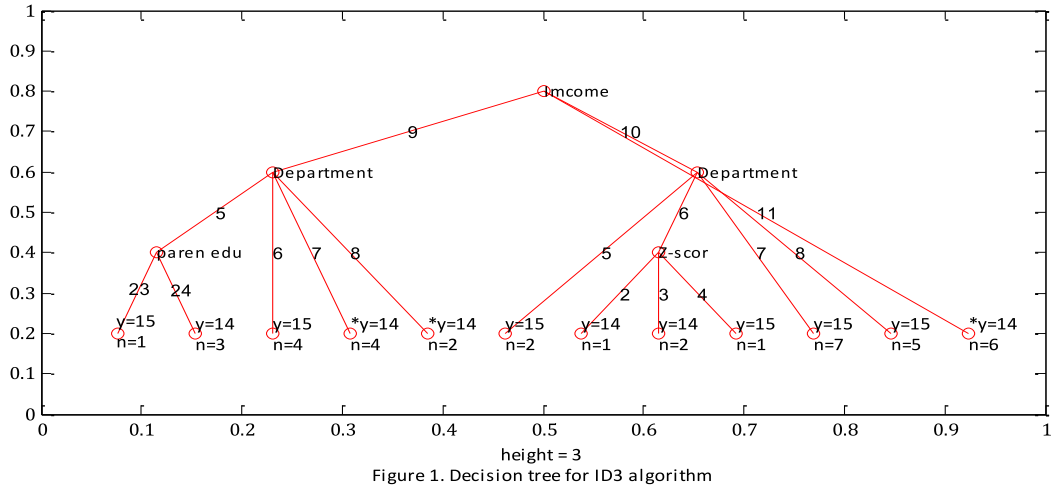


Figure 1. Decision tree for ID3 algorithm

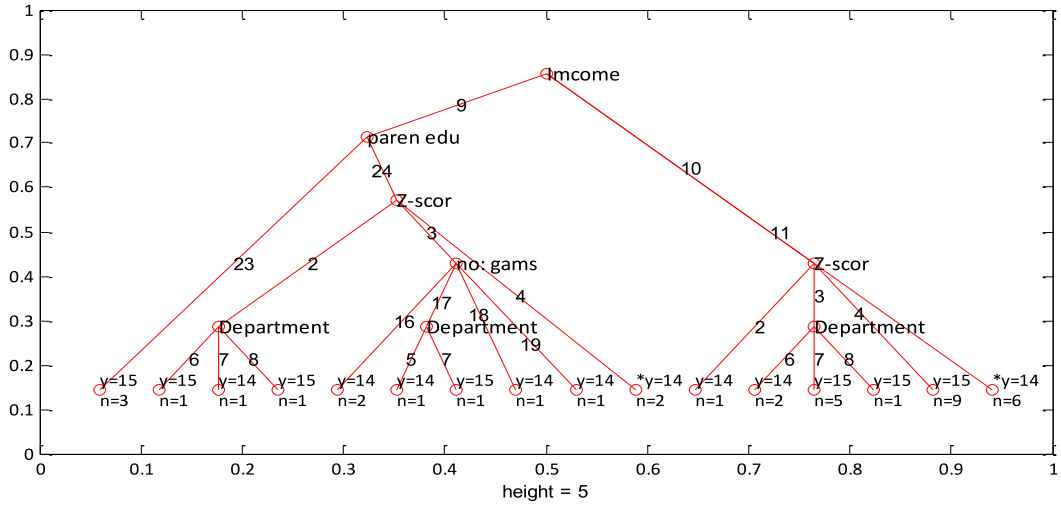


Figure 2. Decision tree for improved ID3 tree algorithm

VIII. CLASSIFICATION ACCURACY

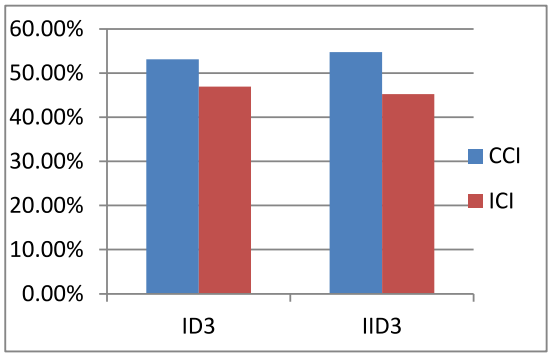
The Table 4 shows the classification accuracy of ID3 and Improved ID3 algorithm on the above data sets.

Table 4: Classifiers Accuracy

Algorithm	Correct Classified instances (CCI)	Incorrect Classified instances (ICI)
ID3	53.0833 %	46.9167 %
IID3	54.9712 %	45.0288 %

It is clear that the correct classification accuracy of the Improved ID3 algorithm is 55.8512% and it is 1.89% improvement of the ID3 algorithm. The classifiers accuracy on previous data sets for each algorithm is represented in the form of a graph.

Figure 3: Comparison of Classification Algorithms



IX. CONCLUSION

In this paper, improved Id3 algorithm is presented to overcome the insufficiency of classical ID3 algorithm. New algorithm makes the decision tree with inescapable increase computational complexity as it needs to compute the values of normalization of relation degree function *V*. But in this work, neglected the computational complexity as the computer technology and the operating speed is rapidly increasing. The improved ID3 algorithm enhances the accuracy of the decision rule than to the classical ID3 algorithm. Further, future works it will be focused on find the prediction accuracy level and apply this improvement with numerical attributes.

ACKNOWLEDGMENT

It is pleasure to acknowledge to the anonymous reviewers of this paper for their helpful suggestions to improve this paper. We would like to thank vice chancellor and the all staff of KDU as well as all students in Intake 28, Faculty of Engineering to give the opportunities to collect the necessary data for this work.

REFERENCES

B.Aashoo, D.Kavita & S.Manish 2012, 'Implementation of Decision Tree', *International Journal of Engineering and Advanced Technology (IJEAT)*, vol II, no. 2, pp. 30-34.

B.Rupali & S.Vatta 2013, 'Implementation of ID3 Algorithm', *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 3, no. Iss 6, pp. 845-851.

C.Jin, L.De-Lin & M.Fen-xiang 2009, 'An Improvised ID3 Decision Tree Algorithm', *Proceedings of 2009 4th International Conference on Computer Science and Education*, pp. 127-130.

Han, J & Kamber, M 2006, *Data Mining: Concepts and Techniques*, 2nd edn, Diane Cerra, Morgan Kaufmann.

K.B.Brijesh & S.Pal 2011, 'Mining Educational Data to Analyze Students" Performance', *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol II, no. 6, pp. 63-69.

M.M.Tom 1997, *Machine Learning*, McGraw-Hill Science/Engineering/Math.

O.D.Samue 2006, 'An Exploration of Classification Prediction Techniques in Data Mining:The insurance domain', Masters Thesis, Bournemouth University, Bournemouth,UK.

S.Anumitha, Diana, S & D.Suganya 2012, 'Improvisation of ID3 Algorithm Explored On Wisconsin Breast Cancer Dataset', *International Conference on Computing and Control Engineering (ICCCE 2012)*.

S.Beniwal & J.Arora 2012, 'Classification and Feature Selection Techniques in Data Mining', *International Journal of Engineering*

Research & Technology (IJERT), vol 1, no. Iss 6.

S.K.Yadav, B.Bharadwaj & S.Pal 2012, 'Data Mining Applications: A comparative Study for Predicting Student's performance', *International Journal Of Innovative Technology & Creative Engineering*, vol 1, no. 12, pp. 13-19.

V.Maduskar & Y.Kelkar 2013, 'A New Modified Decision Tree Algorithm based on ID3', *International Journal of Computer Architecture and Mobility*, vol 1, no. ISS 9, http://www.ijcam.com/index.php- (2014.02.15)

BIOGRAPHY OF AUTHORS

[2]J.K. Wijerathna is a senior Lecturer attached to the department of mathematics, Faculty of Science of the University of Colombo, Sri Lanka. He received his BSc Special degree (first class) at University of Colombo, M.Sc. degree at University of Kaiserslauterm, Germany, and Ph.D Degree from the University of Colombo-Kaiserslautern. His research interests are in Quantitative Finance, Numerical Methods for Partial differential equations, computational Fluid Dynamics and Mathematical Modeling. He has produced several publications to his credit. In addition to his academic qualifications, he was a Consultant Vice President at AMBA Research (Lanka) Ltd. And also he is Life member of Sri Lanka Association for the Advancement of the Science. Currently he is the Act. Dean of the faculty of Science, University of Colombo.

[1]N. S. Rathnayaka is a lecturer attached to the Department of IT and Mathematics, Faculty of Engineering of General Sir John Kothelawala Defence University, Sri Lanka. He completed the BSc Special (Hons) degree in Mathematics with second class (Upper Division) at the University of Ruhuna, Sri Lanka. Currently he is following the Mphil Degree in Mathematics, University of Colombo, Sri Lanka. Rathnayaka's research interests are in Numerical methods, Computational Fluid Dynamics and data mining.