Evaluating the Performance of Teaching Assistant Using Decision Tree ID3 Algorithm

K. Devasenapathy
Information Technology Department
Nehru Arts and Science College
Coimbatore-Tamilnadu -India

S. Duraisamy
Computer Science Department
Chikkanna Government Arts College
Coimbatore-Tamilnadu –India

ABSTRACT
Data mining (DM) is a class of database application that look for the hidden patterns in a collection of data and their relationships. DM is used in developing methods for discovering facts from data which come from educational environment and it becomes educational data mining (EDM). The educational institutions can use classification for complete analysis of students' characteristics. This paper details the Iterative Dichotomiser (ID3) algorithm in classification technique. The ID3 algorithm builds a decision tree from a dataset. This action we accumulates Teaching Assistant Evaluation's (TAE) dataset from UCI machine learning repository. This paper demonstrates the ID3 algorithm to construction of decision tree (DT). The implementation of this algorithm is useful to study of teaching performance over three regular semesters and two summer semesters of 151 Teaching Assistant (TA).In this work various kinds of impurities measures and discovery the maximum information gain at various iterations levels. This task is to extract the knowledge that describes TA performance over summer and regular semester. This exertion will help the institute to growth the performance.

Keywords
Educational Data Mining, Iterative Dichotomiser 3 (ID3) Algorithm, Decision Tree, Teaching Assistant

1. INTRODUCTION
Data Mining (DM) is a gorgeous field of computer science. DM methods applies in the areas of machine learning, statistics, artificial intelligence and databases. DM is the process of automatically searching large stores of data to discover patterns and trends from large data sets. [10] The objective of the DM process is to extract information from a data set and convert it into a logical construction for further use. [5]DM is the process of processing large volumes of data, searching for patterns and relationships within that data. DM has two learning methods like supervised learning and unsupervised learning. Supervised learning is the DM task of understand function from labeled training data**.** The training data consist of a set of training examples. The training data consisting input object and a desired output value. Unsupervised learning is seeking that to find hidden structure in unlabeled data. Supervised learning is the DM task of understand function from labeled test data**.** DM has been applied in numerous fields including e-commerce, bioinformatics, counter terrorism and lately, within the educational research which commonly known as Educational Data Mining (EDM). [2].The EDM society website, www.educationaldatamining.org "an emerging discipline, concerned with developing methods for discovering the unique types of data that come from educational settings, and using those methods to better understanding of the students, and the settings which they learn in". The job of traditional education is to transmit to a next generation. Today the technology development in computer, internet, industrialization and mechanization are blessing of human beings. In the traditional educational system boredom, confusion, engaged concentration, frustration, neutral delight are raised. It must be a transparent transformation which is required from traditional educational system to modern education system [8].

2. EDUCATIONAL DATA MINING (EDM)
The EDM is the application of DM techniques to educational data, and so, its objective is to scrutinize these types of data in order to resolve educational research issues [3].The EDM methods are statistics and visualization, web mining, classification, regression, density estimation, clustering, classification, relationship mining, outlier detection, sequential pattern mining and text mining [6].

The EDM is an emerging field exploring data in educational context by applying different DM techniques and tools. EDM is an interesting research area which extracts useful, previously unknown patterns from educational database for better thoughtful, improved educational performance and assessment of the student learning process .EDM is a term used for processes designed for the analysis of data from educational settings to better to understand students and the settings which they learn in. Today in the EDM there are increasing research interests in using DM techniques in educational field. This new emerging field, EDM, concerns with developing methods that discover knowledge from data which originating from education system. It is often differ from traditional DM techniques. The EDM focuses on cluster, archiving, and analysis of data related to students learning and judgment. The analysis performed in EDM research is often related to techniques drawn from variety of literatures, including psychometrics, machine learning, data mining, educational statistics, information visualization and computational modeling.

3. DECISION TREE
The Decision Tree (DT) is flow-chart tree structure are commonly used for gaining information for the purpose of decision making. Where each inner node is denoted by rectangles and leaf nodes are denoted by ovals. All inner nodes have two or more child nodes. All interior nodes contain splits, which test the value of an expression of the attributes. Arcs from an inner node to its children are labeled with distinct outcomes of the test. Each leaf node has a class label associated with it. DT starts with a root node on which it is for users to take activities. From this node, users split each node recursively according to DT learning algorithm. The final result is a DT in which each branch represents a

possible scenario of decision and its outcome. The Hunt's ID3, CART, CHAID and C4.5 are widely used DT learning algorithms

3.1 Hunt's algorithm is greedy recursive algorithm. This algorithm can use only local optimum on each call without backtracking. Many decision tree induction algorithms are based on this Hunt's algorithm

3.2 ID3 algorithm presented by J. R. Quinlan [7] is a greedy algorithm that selects the succeeding attributes based on the information gain associated with the attributes. The attribute with the maximum information gain or greatest entropy reduction is selected as the test attribute for the recent node.

3.3 C4.5 is an algorithm used to make a decision tree developed by Ross Quinlan. C4.5 is a descendant of ID3. C4.5 made a number of enhancements to ID3. C4.5 uses Gain ratio [4] as an attribute selection measure. Also C4.5 can handle both distinct and continuous attribute

3.4 CART (Classification And Regression Tree) algorithm, which was proposed by Breiman, is conceptually is same as that of ID3. The impurity measure used in choosing the variable in CART is Gini index [4].If the target variable is nominal it produces classification tree and for continuous valued numerical target variable it makes tree.

3.5 CHAID (CHi-squared Automatic Interaction Detector) uses Chi square incident test for tree creation in two ways [1]. First, it determines whether levels in the predictor can be combines together. Once all predictor level is compressed to their smallest weighty form, it governs most important predictor in peculiar among the keep on variable levels.

3.6 The dataset

Teaching Assistant Evaluation **(TAE).**The UCI Machine Learning Repository, Teaching Assistant Evaluation Data Set

was contributed by Wei-Yin Loh and Tjen-Sien Lim (Department of Statistics, University of Wisconsin-Madison).The data consist of evaluations of teaching performance over three regular semesters and two summer

semesters of 151 Teaching Assistant (TA) assignments at the Statistics Department of the University of Wisconsin-Madison. The scores are grouped into three unevenly equal-sized classes (low, medium, and high) to form the class attribute. The predictor attributes are (i) whether or not the TA is a Native English Speaker (NES) (binary), (ii) Course Instructor(CI) (25 categories) (iii) Course categories (26 categories) (iv) Summer or Regular semester (binary) and (v) Class Size(CS) (numerical). It differs from the other datasets in that there are two categorical attributes with large numbers of categories

4. THE ID3 DECISION TREE

ID3 stands for induction decision tree-version 3.In ID3,a recursive procedure is used to construct a decision tree from data [9].The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy examination through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for categorizing a given sets, introducing a metric - information gain to find an optimal way to organize a learning set, what we essential to do is to curtail the questions asked (i.e. minimizing the depth of the tree). According, we need some functions which can degree, which questions provide the most balanced splitting.

4.1 Measuring Impurity

Given a data table that contains attributes and class of the Attributes are measure homogeneity (or heterogeneity) of the table based on the classes. A table is pure or homogenous if it encompasses only a single class. If a data table contains several classes, then that the table is impure or heterogeneous. There are several indices to measure degree of impurity quantitatively. Utmost fine recognized indices to produce degree of impurity are entropy, gini index, and classification error.

$$Probability(Low) = \frac{52}{151} = 0.344$$

$$Probability(Medium) = \frac{50}{151} = 0.331$$

$$Probability(High) = \frac{49}{151} = 0.325$$

Entropy

Entropy is way measure the impurity. It is calculated based on proportion of target values. The formula as follows

$$H(X) = \sum_{j}^{n} -p_j \log_2 p_j$$

$$The\ logaritham\ base\ is\ 2$$

$$Entropy = -0.344 \log_2 0.344 - 0.331 \log_2 0.331 - 0.325 \log_2 0.325 = 1.585$$

4.2 Information Gain

The information gain metric is such a function

$$For\ Set\ S, Attribute\ A\ Where\ S\ is\ split\ into$$

$$subsets\ based\ on\ values\ of\ A \quad \subset_S^A = Subset\ A\ of\ S$$

$$I_{E\ =Entropy} \quad \subset_S^A x = \frac{Size(\subset_S^A)}{\text{Size } (S)}$$

$$I_{G(S,A)\ =}\ I_E\ (S) - \sum^n ((p(\subset_S^{A_n}) * I_E(\subset_S^{A_n}))$$

Results of First Iteration

Table 1: Entropy based maximum information gain - Class Size (CS)

Gain	Whether of TA is a NES / Non-NS	Course Instructor	Course	Summer/ Regular	Class Size	Maximum Information Gain
Entropy	0.056	0.405	0.434	0.062	0.820	0.820

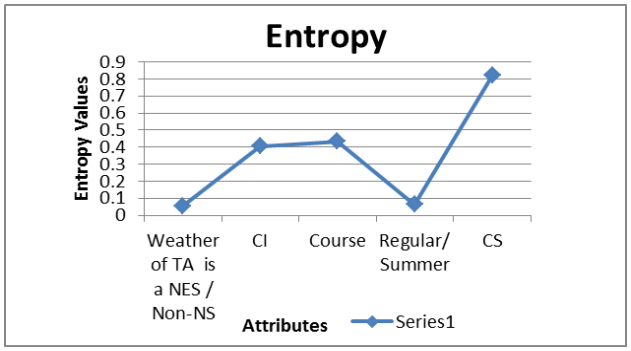

Fig 1: Result of first iteration Class Size (CS) is maximum information gain

The above Fig 1 explains the entropy based maximum information gain. It indicates the maximum information gain is Class Size (CS). The Class Size becomes the root node of the DT.

Results of Second Iteration

TABLE 2: Entropy based maximum information gain - Course Instructor (CI)

Gain	Whether of TA is a NES / Non-NS	Course Instructor	Course	Summer/ Regular	Maximum Information Gain
Entropy	0.414	0.581	0.538	0.103	0.585

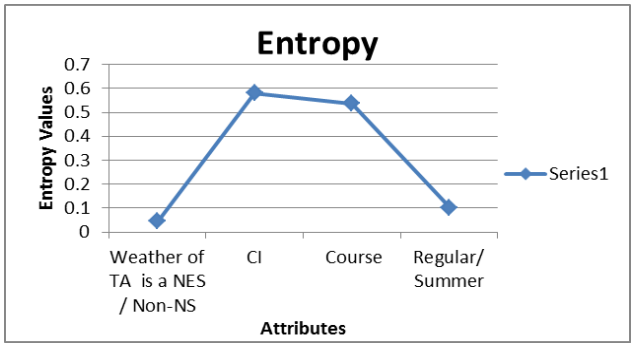

Fig 2: Result of second iteration Course Instructor (CI) is maximum information gain

The above Fig 2 explains the entropy based maximum information gain. It indicates the maximum information gain is Course Instructor (CI). The Course Instructor becomes the next level of the Decision Tree (DT)

Results of Third Iteration

Table 3: Entropy based maximum information gain – Course

Gain	Whether of TA is a NES / Non-NS	Course	Summer/ Regular	Maximum Information Gain
Entropy	0.083	0.540	0.106	0.540

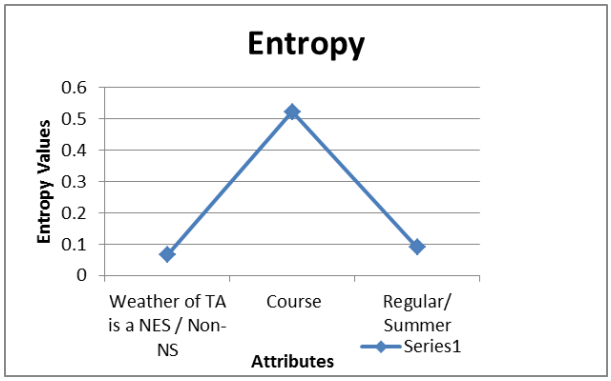

Fig 3: Result of third iteration Course is maximum information gain

The above figure 3 explains the entropy based maximum information gain. It indicates the maximum information gain is Course. The Course becomes the next level of the DT.

Results of Fourth Iteration

TABLE 4: Entropy based maximum information gain - TA is a NES/Non -NS and Summer/ Regular

Gain	Whether of TA is a NES / Non-NS	Summer/ Regular	Maximum Information Gain
Entropy	0.073	0.143	0.143

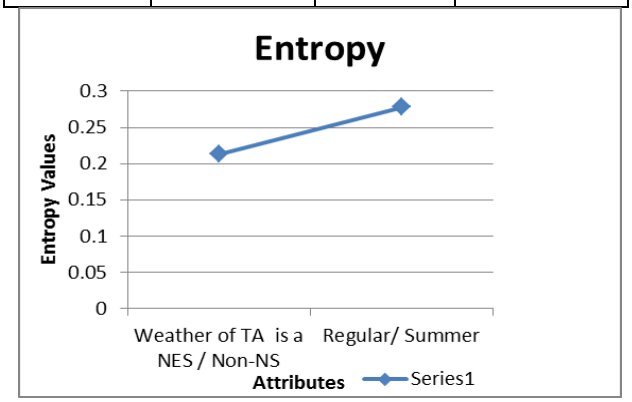

Fig 4: Result of fourth iteration TA is a NES/Non-NS and Regular/Summer

The above figure 4 explains the entropy based maximum information gain. It indicates the maximum information gain is equal to reaming attributes of whether of Teaching Assistant is a Native English Speaker (NES) or Non-Native Speaker (NS) and Regular or Summer. Both are of equal level in the DT.

5. CONSTRUCTION OF DECISION TREE USING ID 3 ALGORITHM

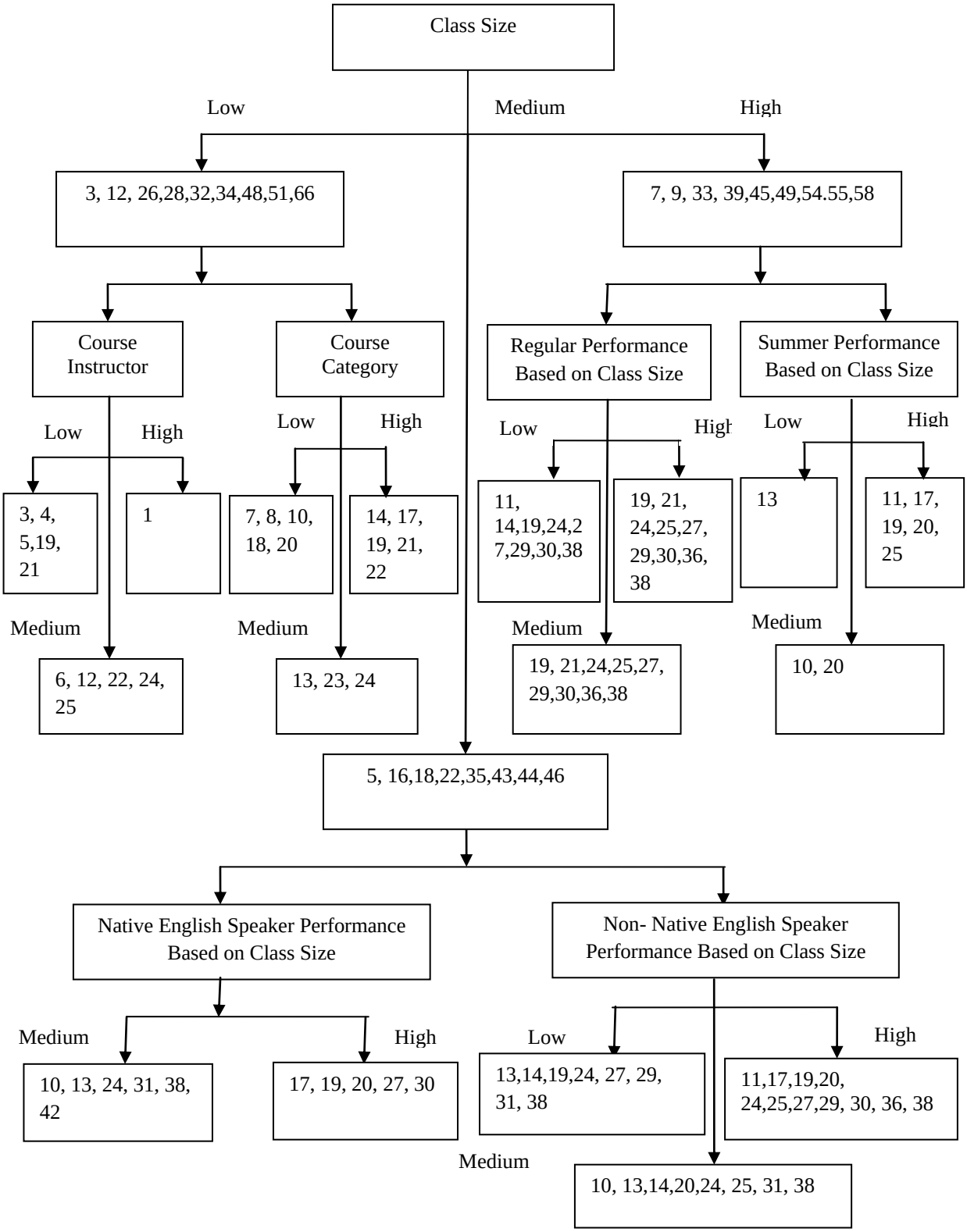

FIGURE 5: **Construction of Decision Tree for Teaching Assistant Evaluation dataset using ID3 Algorithm**

6. CONCLUSION

The DM methods are applied in education field it is called EDM. The EDM techniques are used to improve the process of educational settings like Schools, Colleges and Universities. EDM is an application of DM. The EDM is promising research field. In this paper, implementation algorithm of ID3 algorithm in the Teaching Assistant Evaluation (TAE) dataset and the approach of decision tree (DT) induction using ID3 algorithm. ID3 algorithm calculated different types of impurities and finding the maximum information gain at various levels of iteration. It helps to analysis the performance of teaching assistant (TAE) evaluation dataset with different dimensions. The DT shows the overall performance of the TA. The Experimental results are hopeful, investigating other relational dataset. The implementation of classification algorithm in educational data set, retriever the improved result in future works.

7. REFERENCES

[1] Alaa el-Halees, "Mining students data to analyze eLearning behavior: A Case Study", 2009.

[2]Baker, R.S.J.d. (2010) Data Mining for Education. In. McGaw, B., Peterson, P., Baker, E. (Eds.) International. Encyclopedia of Education 3rd edition. Elsevier, Oxford (2010).

[3] Barnes, T., Desmarais, M., Romero, C., Ventura, S. Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings. Cordoba, Spain.

[4] Bharadwaj B.K. and Pal S. "Mining Educational Data to Analyze Students‟ Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69, 2011.

[5] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[6] Expert Systems with Applications : Educational data mining*:* A survey from 1995 to 2005 Volume 33, Issue 1, July 2007, pages 135–146.

[7] Quinlan, J. R. *(*1986*).* "Introduction of decision tree", Machine learn, 1*:* pp. 86-106.

[8] Devasenapathy.K and Duraisamy "Foreword of Computer Based Techniques in Educational Data Mining and Applying Data Mining Methods in Traditional Educational System", International Journal Applied Engineering Research ISSN 0973- 4562 Vol.10 No.21 pages 20375–20381.

[9] Soman K.P, Shyam Diwakar and Ajay.V ", Insight into Data mining Theory and Practice", Easter Economy Edition, Prentice Hall of India, 2006, ISBN-81-203-2897-3 Pages 403

[10] U.S. Department of Education, Office of Educational Technology, Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief, Washington, D.C., 2012.