Asian Journal of Information Technology 15 (19): 3820-3825, 2016
ISSN: 1682-3915

Performance Analysis of Teaching Assistant Using Decision Tree Classification Algorithm

[1]K. Devasenapathy and [2]S. Duraisamy
[1]Department of Computer Applications, Tamil Nadu College of Engineering, 641008 Tamil Nadu, India
[2]Department of Computer Applications, Sri Sri Krishna College of Engineering and Technology, 641008 Tamil Nadu, India

Abstract: Data mining is a powerful tool for pedagogic intervention. When data mining is used in developing methods for discovering knowledge 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 study details the Hunt's algorithm in classification technique. The Hunt's algorithm builds a decision tree from a dataset. In our work we collect Teaching Assistant Evaluation's (TAE) dataset from UCI machine learning repository. Then we applied Hunt's Algorithm to construction of Decision Tree (DT). The implementation of this algorithm is useful to analysis of evaluations of teaching performance over three regular semesters and two summer semesters of 151 Teaching Assistant (TA).By this work we find various types of impurities measures and finding 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 work will help the institute to improve the performance.

Key word: Educational data mining, classification, hunt's algorithm, decision tree, teaching assistant

INTRODUCTION

Data Mining (DM) is a set of database application that looks for the hidden patterns in a group of data and their relationships. DM is the process of automatically searching large stores of data to find out model and affinity from huge data sets. The facts and knowledge is learning from data. This learning from DM comes in two outlines: that is supervised learning and unsupervised learning. In supervised learning or directed DM mining, it explains the data by using the target field of the dataset .The unsupervised learning or undirected DM, it tells data by using without the target field of the dataset. There is unanimity that DM is not a single-step process and knowledge discovery is the result of successive processes The process of DM is also known as Knowledge Discovery in Data (KDD). 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) (Baker, 2010).

MATERIALS AND METHODS

Educational Data Mining (EDM): Education is for a sustainable living, students can gain knowledge, skills and values to address the environmental and social challenge of the coming decades (center for ecoliteracy, in 2009). The purpose of education should be to develop a love of learning that continues with students with during their lives (Education and Motivation By Captain Bob Webb).The business 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 (Devasenapathy and Duraisamy, 2008).

The EDM is the application of DM techniques to educational data and so, its objective is to analyze these types of data in order to determine educational research issues (Baker and Yacef, 2009). The EDM society website, www.educationaldatamining.org "an rising discipline, concerned with increasing methods for exploring 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 EDM methods are statistics and visualization, web mining,

Corresponding Author: K. Devasenapathy, Department of Computer Applications, Tamil Nadu College of Engineering, 641008 Tamil Nadu, India

classification, regression, density estimation, clustering, classification, relationship mining, outlier detection, sequential pattern mining and text mining (Romero and Ventura, 2010). The EDM is a rising exploring data in educational context by applying different data mining techniques and tools. EDM is an interesting research area which extracts helpful, previously unidentified patterns from educational database for better understanding, improved educational performance and assessment of the student learning process (Chan *et al*., 2008). 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 developing field, EDM, concerns with developing methods that determine knowledge from data which originating from education system. It is often differ from traditional data mining techniques. The EDM focuses on collection, archiving and analysis of data related to students learning and evaluation. 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.

EDM methods: The types of EDM method are related to those found in data mining in general, but with some differences based on the unique features of educational data. The EDM classifies into prediction, clustering, relationship mining and discovery with models (Baker and Yacef, 2009).The discovery of with models, a model of a experience is developed during process that can be validated in some method this model is then used as a component in another analysis. The following are primary application of EDM are analysis and visualization of data, feedback for supporting instructors, recommendations for students, predicting student performance, student modeling, detecting undesirable student behaviors, grouping students, social network analysis, developing concept maps, constructing courseware and planning and scheduling (Romero and Ventura, 2010).

M-classification method: Databases are loaded with covered information that can be used for intelligent decision making. Classification is a data mining (machine learning) technique that assigns items in a collection to target categories or classes. It is a supervised learning method which requires labeled training data to produced rules for classifying test data into predetermined groups or classes (Dunham, 2006). Classification and prediction is two sketch of data analysis that can be used to extract models describing imperative data classes or to predict future data trends (Vaishnavi, 1989). Prediction has implicated considerable attention given the potential implications of successful forecasting in business context. Classification is a two-phase process. The first phase is the learning phase, where the training data is analyzed and grouping rules are generated. The next phase is the classification, where test data is arranged into classes according to the generated rules. The goal of classification is to accurately predict the intent class for each case in the data. For example, a classification model could be used to identify teaching performance as low, medium or high. Such analysis can help provide us with a better concerned of the data at large. The classification methods are decision Tree based methods, rule-based methods, neural networks, genetic algorithms, naïve byes and bayesian belief networks and support vector machines.

Decision tree: A Decision Tree (DT) is a flow-chart tree structure are commonly used for gaining information for the purpose of decision making. DT is a very popular approach and practical approach for a pattern classification. DT is a greedy, top-down and recursive process. The DT method is a supervised machine learning technique that builds a DT from a set of class labeled training samples during the machine learning process (Han *et al*., 2011). DT classifies data into discrete ones using tree structure algorithms (Qinlan, 1986). The main purposes of DT's are to represent the structural information contained in the data. In DT induction we are used Hunts algorithm, CART, ID3, C4.5, SLIQ and SPRINT.

The dataset: Teaching Assistant Evaluation (TAE). The UCI dataset 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 roughly equal-sized categories (low, medium and high) to form the class attribute. The predictor attributes are whether or not the TA is a Native English Speaker (NES) (binary), Course Instructor (CI) (25 categories) Course (26 categories) (iv) Summer or Regular semester (binary) and Class Size (CS) (numerical). It differs from the other datasets in that there are two categorical attributes with large numbers of categories (UCI Machine Learning Repository, Teaching Assistant Evaluation Data Sets).

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

Pseudocode for hunts algorithms: Build Tree (Node t, Training Database D, Split Selection Method S)

- Apply S to D to find splitting criterion
- If (t is not a leaf node)
- Create children nodes of t
- Partition D into children partitions
- Recursive on each partition
- Endif

Measuring impurity: From our dataset (parent Table 1) Contains attributes and class of attributes. We measure the impurity of the parent table D. They are well known indices measures the impurity, they are entropy, gini index and classification error.

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

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

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

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

$$\text{Entropy}(S) = \sum_{j}^{n} -p_j \log_2 \log_2 p_j$$

The logaritham base is 2

$$\text{Entropy} = -0.344\log_2 0.344 - 0.331\log_2 0.331$$
$$-0.325\log_2 0.325 = 1.585$$

Gini Index: Gini Index is another way to measure impurity. The formula as follows

$$\text{GiniIndex}(G) = 1 - \sum_{j} p_j^2$$

$$\text{GiniIndex}(G) = 1 - (0.344^2 + 0.331^2 + 0.3325^2)$$
$$= 0.666$$

Table 1: Impurity values of parent table

Impurity measure	Entropy	Gini index	Classification error
Values	1.585	0.666	0.656

Table 2: Entropy based maximum information gain - Class Size (CS)

Gain	Whether of TA is a NES/Non-NS	Course instructor	Course	Summer/ Regular	Maximum information gain
Entropy	0.056	0.405	0.434	0.062	0.820

Table 3: 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	0.585	0.535	0.103	0.585

Classification error: Classification error is another way to measure impurity. The formula as follows:

$$\text{ClassificationError}(T) = 1 - \max\{p_j\}$$

$$\text{ClassificationError}(T)$$
$$= 1 - \max\{0.344, 0.331, 0.325\}$$
$$= 1 - 0.344 = 0.656$$

The value of probability is p^j is a class j

Information gain: The information gain is measure of this change in entropy. It has computed impurity degree between parent table and subset table. The measure and compare the different impurity degree is called information gain (Table 2 and 3):

$$\text{Gain Information Gain }(A)$$
$$= \text{Entropy}(S) - \sum_{i=1}^{k} \left|\frac{Si}{S}\right| \text{Empty}(S_i)$$

RESULTS AND DISCUSSION

Results of first iteration: The Fig. 1 and Table 3 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 Decision Tree (DT).

Results of second iteration: Figure 2 and Table 4 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: Figure 3 and Table 5 explains the entropy based maximum information gain. It indicates

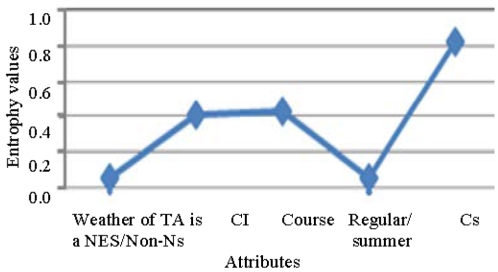

Fig. 1: Result of first iteration-Class Size (CS) is maximum information gain

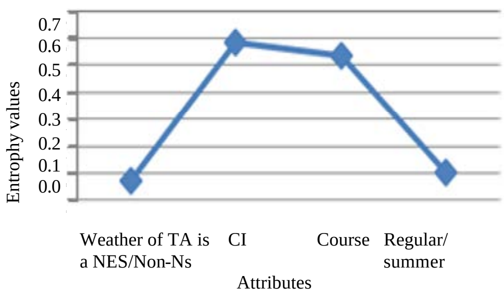

Fig. 2: Result of second iteration-Course Instructor (CI) is maximum information gain

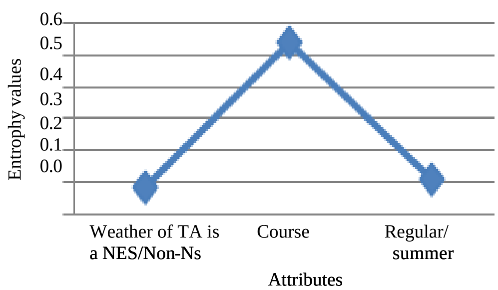

Fig. 3: Result of third iteration course is maximum information gain

the maximum information Gain is course. The course becomes the next level of the Decision Tree (DT)

Results of fourth iteration: 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 of are equal level in the Decision Tree (DT)

Construction of decision tree using hunts algorithm: From Fig. 5 shows how the Hunts algorithm works. We have data records that contain attributes that are associated with classes: We calculate information gain based on entropy and find maximum one that will become becomes the root node of the DT. The maximum information indices entropy is zero that attribute will become the leaf node of the DT. Further iteration the maximum information attribute is eliminated Branches are grown based each outcome of information gain. From the first iteration class size becomes the maximum information gain so that 45 class sizes are becomes purity. Out of 45 class sizes interesting one point is consider the class size numerical attribute 51.The 51 class size is appeared in three times in TAE data set and categorical attribute is also low. There is no crossing between the categorical attribute of low, medium and high. During the second iteration course instructor is a maximum information gain, totally 22 Course Instructor are eliminated. During the third iteration course is a maximum information gain, totally 16 course are eliminated. Based on the categorical attribute of class we split into decision tree levels low, medium and high.

The maximum information gain (0.143) is equal of the predictor attributes of whether or not the TA is a Native English Speaker (14) (binary) and summer or regular semester (binary). From above decision tree figure the Native English Speaker scores low performance attributes is zero, even the medium and high are equal. Native English Speaker is slightly giving the strengthen. And also there is no big difference between the Non-Native English (52) scores are equal (low, medium, high). The

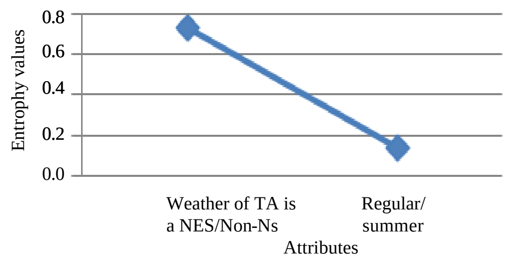

Fig. 4: Result of fourth iteration TA is a nes/non-ns and regular/summer is equal

Table 4: 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

Table 5: Entropy based maximum information gain -TA is a NES/Non -NS and Summer or Regular

Gain	Whether of TA is a NES/Non-NS	Summer/ regular	Maximum information gain
Entropy	0.073	0.143	0.143

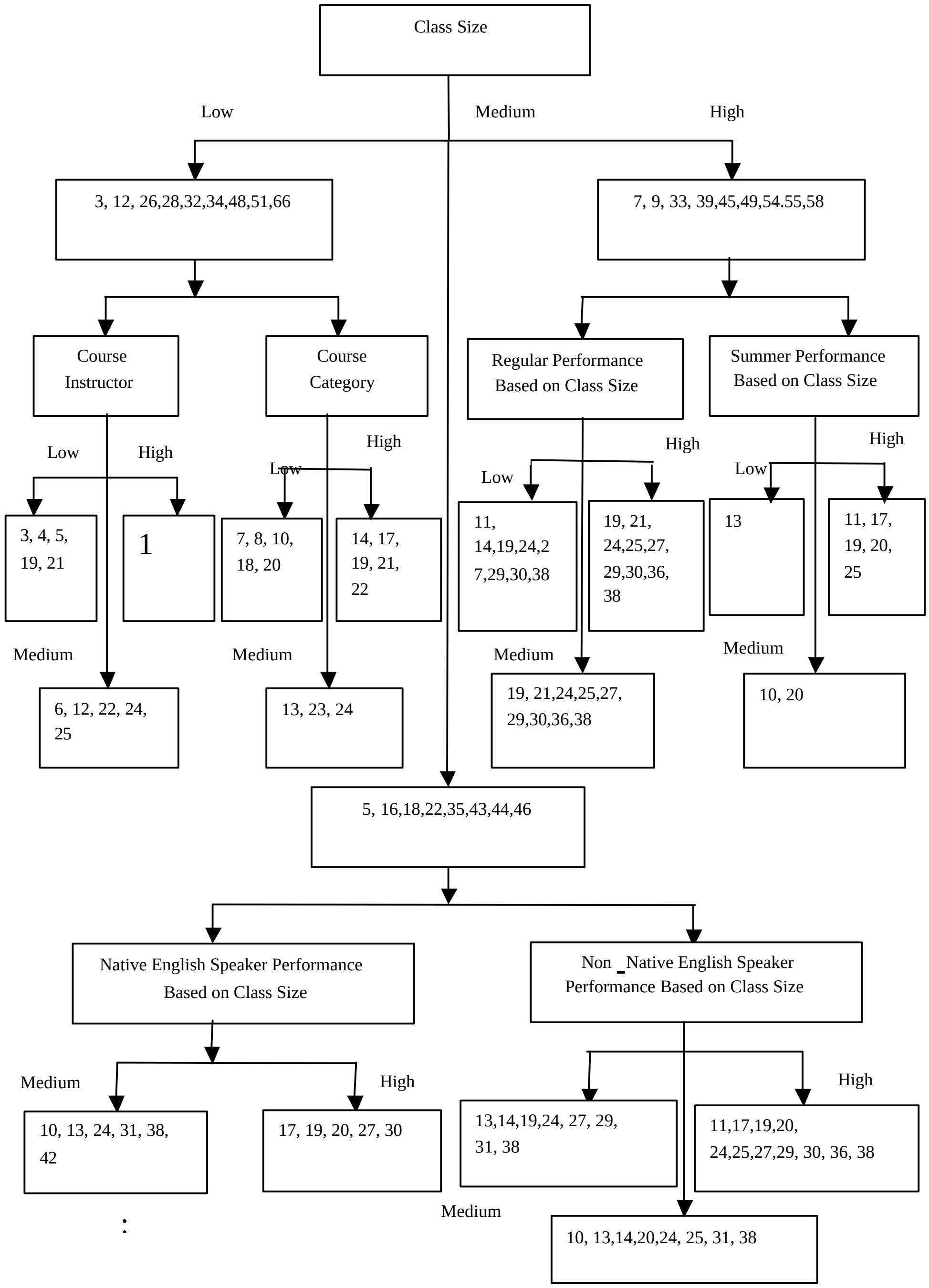

Fig. 5: Construction of decision tree for teaching assistant evaluation dataset

regular and summer semester are remaining attributes. From that the summer semester performance TA is only two attributers.

CONCLUSION

Educational Data Mining could be used to improve the process of educational settings like Schools, Colleges and Universities. EDM is an application of DM, statistics and a promising research field. In this study we implements algorithm of Hunt's algorithm in the Teaching Assistant Evaluation (TAE) dataset. And we have seen the approach of Decision Tree (DT) induction using Hunts algorithm. We 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 encouraging, analyzing other relational dataset. The classification techniques suggested to achieve better results in future works.

REFERENCES

Baker, R.S. and K. Yacef, 2009. The state of educational data mining in 2009: A review and future visions. JEDM-J. Educ. Data Mining, 1: 3-17.

Baker, R.S.J.D., 2010. Data mining for education. Int. Encycl. Educ., 7: 112-118.

Chan, A.Y.K., K.O. Chow and K.S. Cheung, 2008. Online course refinement through association rule mining. J. Educ. Technol. Syst., 36: 2007-2008.

Devasenapathy, K. and D.S. Duraisamy, 2008. Foreword of computer based techniques in educational data mining and applying data mining methods in traditional educational system. Int. J. Appl. Eng. Res., 10: 20375-20381.

Dunham, M.H., 2006. Data Mining: Introductory and Advanced Topics. Pearson Education India, New Dehli, India, ISBN: 978-81-7758-785-2, Pages: 301.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA., ISBN-13: 9780123814791, Pages: 744.

Qinlan, J.R., 1986. Introduction of Decision Trees. Mach. Learn., 1: 81-106.

Romero, C. and S. Ventura, 2010. Educational data mining: A review of the state of the art. IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev., 40: 601-618.

Vaishnavi, V.K., 1989. Multidimensional balanced binary trees. IEEE. Trans. Comput., 38: 968-985.