ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.2

A Comparative Analysis of Student's Continuous Internal Assessment using Decision Tree Algorithms

K. Uma Maheswari[1], Greeshma[2]
Assistant Professor[1], PG Student[2]
Department of MCA
Karpagam College of Engineering, Tamil Nadu, India

Abstract:
KDD process is used for making decisions in educational database system. A decision tree classifier is a machine learning technique used for data exploration. This paper infers the use of decision trees in educational data mining. Decision tree algorithms are applied on students' continuous internal performance data to generate the model and this model is used to predict the students' Continuous Internal Assessment performance. This model helps the teachers in earlier to identify the students who need special attention and appropriate advising/counseling.

Keywords: Educational Data Mining, Classification, Knowledge Discovery in Database (KDD), Decision Tree, Machine Learning.

I. INTRODUCTION

Students are main assets of universities/ Institutions. The student's academic performance plays a vital role in producing the superlative graduates and post-graduates who will become great leader and manpower for the country. The performance of students in universities should be a concern to corporations in the market. Academic achievement is one of the main factors considered by the employer in recruiting workers especially the fresh graduates. Thus, students have to place the greatest effort in their study to obtain a good grade in order to fulfill the employer's demand. Academic achievement of a particular student is considered by their Cumulative Grade Point Average (CGPA) mark. CGPA shows the overall student's academic performance where it considers the average of all examinations' grade for all semesters during the tenure in university. Several factors could act as hurdle and medium to students achieving a high CGPA that reflects their overall academic performance. Since, the internal mark of the student plays an important role. The data collected from various applications require appropriate method for extracting knowledge from huge repositories for better decision making. Knowledge discovery in databases (KDD) aims to discover the useful information from volume of data [1]. The main functions of data mining are applying a variety of methods and algorithms in order to discover and extract patterns from data storage [1]. Data mining tools predict patterns, future trends and behaviors, allowing businesses to effect proactive, knowledge-driven decisions. Educational Data Mining is a promising field, having superior methods to determine knowledge from the data emanating from the databases related to educational environments [2]. There are many techniques used in Educational Data Mining such as Decision Trees, Neural Networks, Naïve Bayes, K-Nearest neighbour, and many others [1]. The main aim of this paper is to use data mining methodologies to study students' continuous internal performance in the courses. Data mining provides many tasks that could be used to study the student's continuous internal assessment performance. In this research, the classification task is used to evaluate student's continuous internal assessment performance and as there are many approaches that are used for data classification, the decision tree method is used here. Student's information likes CIA I, CIA II and CIA III marks were collected from the student's management system, to predict the performance at the end of each CIA's examination.

II. DECISION TREE INTRODUCTION

Decision tree is like a complete binary tree, where each internal node is represented by rectangles, and leaf nodes are represented by ovals. All internal nodes have two or more children. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labeled with different outcomes of the test. Each leaf node has a class label associated with it.

The decision tree classifier has two phases:

i) Tree construction phase.
ii) Tree Pruning phase.

In the tree construction phase, decision tree is built based on the training data set by recursively splitting the training set until all or most of the records belonging to each of the partitions belonging to the same class label. The prune phase generalizes the tree by removing the noise and outliers. This phase increases the accuracy of the classification.

Table.1. Frequency Usage of Decision Tree Algorithms

Algorithm	Usage frequency (%)
CLS	10
ID3	67
IDE3+	4.5
C4.5	54.55
C5.0	9
CART	40.9
Random Tree	4.5
Random Forest	10
SLIQ	27.27
Public	13.6
OCI	4.5
Clouds	4.5

Tree pruning phase accesses only the fully grown tree. The tree construction phase requires multiple passes over the training data. The time needed for pruning the decision tree is very less compared to build the decision tree. The table I specified represents the usage frequency of various decision tree algorithms [17]. Observing the above table the most frequently used decision tree algorithms are ID3, C4.5 and CART. Hence, the experiments are conducted on the above three algorithms.

A. **ID3 (Iterative Dichotomiser 3)**

This is a decision tree algorithm introduced in 1986 by Quinlan Ross [4]. It is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used. Information gains is calculated for all attribute and choose the attribute with the highest information gain as a root node for building a decision tree. Label the attribute as a root node and the possible values of the attribute are represented as arcs. All possible outcomes are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances. ID3 can be used for continuous attributes, to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

B. C4.5:

C4.5 is an algorithm used for generating a decision tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. It can be used for classification. C4.5 is also referred to as a statistical classifier. C4.5 divides the attribute values into two partitions based on the threshold to handle the continuous attributes such that all the values above the threshold is considered as one child and the remaining as another child. It handles missing attribute values also. For building a decision tree C4.5 uses Gain Ratio as an attribute selection measure. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

C. ***CART:*** CART stands for Classification and Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. Gini Index is one of the attribute selection measures to build a decision tree used by CART. Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it discovers binary trees. ID3, C4.5 uses probabilistic assumptions but Gini Index measure does not use that. Cost complexity pruning is applied to remove the unreliable branches from the decision tree to improve the accuracy.

III. DATA MINING PROCESS:

In present day's educational system, a student's performance is determined by the internal assessment and end semester examination. The internal assessment is carried out by the teacher based upon student's performance in educational activities such as class test, seminar, assignments, general proficiency, attendance and lab work. The end semester examination is one that is scored by the student in semester examination. Each student has to get minimum marks to pass a semester in internal as well as end semester examination.

A. Data Preparations

The data set used in this study was obtained from the sampling method of computer Applications department of course MCA (Master of Computer Applications) from session 2016 to 2017. Initially size of the data is 46. In this step data stored in different tables was joined in a single table.

B. Data Selection and Transformation

In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table II for reference.

The domain values for some of the variables were defined for the present investigation as follows:

CIA – Continuous Internal Assessment, Here in each semester three internal tests are conducted and average of three internal tests are used to calculate average marks. CIA is split into ten classes like range1, range2, etc.

AVERAGE – Average is obtained by calculating the average marks of all CIA's. It is split into two classes: Pass >=50% and Fail < 50%.

C. Data Set: The data set of 32 students used in this study was obtained from Computer Applications department of course MCA (Master of Computer Applications) from session 2016 to 2018.

D. Model Construction

The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the Weka software.

Table .2. Students Related Variables

Variable	Description	Possible Values

CIA I	Continuous Internal Assessment I	{
		range1:0-10,
		range2:11-20,
		range3: 21-30,
		range4: 31-40,
		range5: 41-50,
		range6:51-60,
		range7:61-70,
		range8:71-80,
		range9:81-90,
		range10:91-100
		}

CIA II	Continuous Internal Assessment II	{
		range1:0-10,
		range2:11-20,
		range3: 21-30,
		range4: 31-40,
		range5: 41-50,
		range6:51-60,
		range7:61-70,
		range8:71-80,
		range9:81-90,
		range10:91-100
		}

CIA III	Continuous Internal Assessment III	{
		range1:0-10,
		range2:11-20,
		range3: 21-30,
		range4: 31-40,
		range5: 41-50,
		range6:51-60,
		range7:61-70,
		range8:71-80,
		range9:81-90,
		range10:91-100
		}

AVERAGE	Average of CIA I,II & III	{pass >= 50%, Fail<50%}

It is a machine learning software to solve data mining problems. Weka tool supports various data mining tasks such as data preprocessing, clustering, classification and regression etc. It provides access to SQL databases and also deep learning. From the given data, CIA_MCA.arff file was created. This file was loaded into WEKA explorer. Using this weka explorer, user can apply classification and regression algorithms to the resulting dataset, to find the accuracy. There are 16 decision tree algorithms like ID3, J48, CART etc. implemented in WEKA. The algorithm used for classification is ID3, C4.5 and CART.

Table.3. Data Set

S.NO	CIA I	CIA II	CIA III	AVERAGE
1	range3	range4	range4	FAIL
2	range4	range5	range5	FAIL
3	range2	range5	range5	FAIL
4	range2	range6	range6	FAIL
5	range3	range7	range7	PASS
6	range6	range6	range6	PASS
7	range7	range7	range7	PASS
8	range3	range3	range8	PASS
9	range8	range8	range9	PASS
10	range3	range3	range5	FAIL
11	range8	range8	range1	PASS
12	range3	range3	range2	FAIL
13	range6	range6	range6	PASS
14	range3	range3	range7	FAIL
15	range8	range8	range7	PASS
16	range6	range6	range6	PASS
17	range6	range6	range8	PASS
18	range7	range8	range9	PASS
19	range9	range9	range9	PASS
20	range8	range8	range8	PASS
21	range9	range9	range9	PASS
22	range5	range5	range5	PASS
23	range1	range1	range1	FAIL
24	range2	range2	range2	FAIL
25	range6	range6	range6	PASS
26	range7	range7	range7	PASS
27	range7	range7	range7	PASS
28	range6	range6	range6	PASS
29	range6	range6	range6	PASS
30	range7	range7	range7	PASS
31	range7	range7	range7	PASS
32	range6	range6	range6	PASS

E. Results Obtained

The accuracy of ID3, C4.5 and CART algorithms applied on the above data sets shown in table III and the classifiers accuracy is shown in the table IV.

Table.4. Classifiers Accuracy

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
ID3	55.0834%	44.9166%
C4.5	47.8334%	52.1666 %
CART	58.50%	41.50%

From the Table IV, CART technique has highest accuracy of 58.50% compared to other techniques used on same data set. ID3 algorithm is also showing an acceptable level of accuracy. The classifiers accuracy of various methods on the above data sets is represented in the form of a graph.

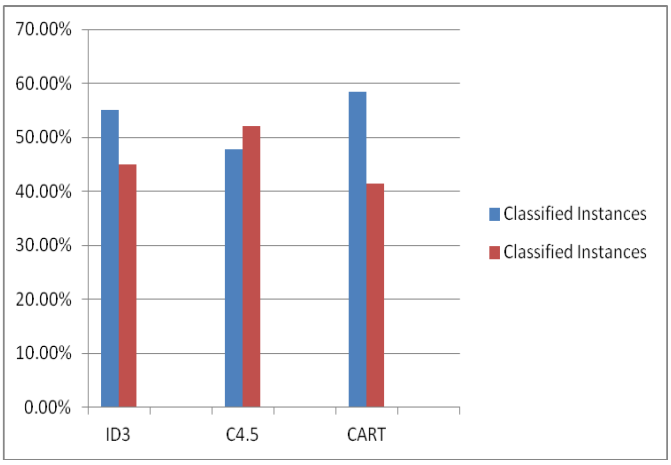

Figure.1. Comparison of Classifiers

IV. CONCLUSION

Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are well-liked because it produces classification rules that are interpreted easily than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Student data to predict the student's performance in the end semester examination. The experimental results show that CART is the best algorithm for classification of data. This model will help the students and the teachers to improve the performance of the students. This model will also work to identify the student those who needed special attention and will also work to reduce fail ratio and taking appropriate action to improve student's performance in the next continuous internal assessments and end semester examination.

V. REFERENCES

[1]. Ms. K. Uma Maheswari, Mrs. R. Gokila , Dr. S. Angel Latha Mary "Mining Educational Data for Predicting Students Subject Knowledge Using Decision Support System", International Journal of Pure and Applied Mathematics, Vol.118, pp. 307-311, 2018.

[2]. Ms. K. Uma Maheswari, Mrs. R. Gokila, Mr. K. Senthil Kumar," Evaluating Student's Eligibility For Recruitment Process Using Classification", Karpagam Journal of Engineering and Research,2017.

[3]. Brijesh Kumar Hardwar," Data Mining: A prediction for performance improvement using classification", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2014.

[4]. Dr. V. Narayani G. Dona Rashmi, K. Uma Maheswari, "Analytical Research in the Geographical Area for Classifying Childhood Obesity Using ID3 Algorithm", International Journal of Computer Science, Vol. 2, 2014.

[5]. Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal," Data Mining Applications: A comparative Study for Predicting Student's performance ", International Journal of Innovative Technology & Creative Engineering, Vol.1, No.12, ISSN: 2045-711, 2014.

[6]. U. K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN: 0975-9646, 2011.

[7]. U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN: 1694-0814, 2011.

[8]. B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

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

[10]. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[11]. Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference 2010.