Data Mining: A prediction for Student's Performance Using Decision Tree ID3 Method

D.BHU LAKSHMI, S. ARUNDATHI, DR.JAGADEESH

Abstract**—** Knowledge Discovery and Data Mining (KDD) is a multidisciplinary area focusing upon methodologies for extracting useful knowledge from data and there are several useful KDD tools to extracting the knowledge. This knowledge can be used to increase the quality of education. But educational institution does not use any knowledge discovery process approach on these data. Data mining can be used for decision making in educational system. A decision tree classifier is one of the most widely used supervised learning methods used for data exploration based on divide & conquer technique. This paper discusses use of decision trees in educational data mining. Decision tree algorithms are applied on students' past performance data to generate the model and this model can be used to predict the students' performance. The most useful data mining techniques in educational database is classification, the decision tree (ID3) method is used here**.**

Index Terms— Educational Data Mining, Classification, Knowledge Discovery in Database (KDD), ID3 Algorithm.

IJSER

1. Introduction

The advent of information technology in various fields has lead the large volumes of data storage in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1]. The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [2]. The main objective of this paper is to use data mining methodologies to study student's performance in end General appreciation. Data mining provides many tasks that could be used to study the student performance. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here.

2. Related Work

Han and Kamber (1996) [3] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

Brijesh Kumar Baradwaj and Saurabh Pal (2011) [1] describes the main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. Alaa El-Halees (2009) [4] applied the educational data mining concerns with developing methods for discovering knowledge from data that come from educational environment. used educational data mining to analyze learning behavior. Student's data has been collected from Database course. After preprocessing the data, we applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, we extracted knowledge that describes students' behavior.

Mohammed M. Abu Tair and Alaa M. El-Halees (2012) [5] applied the educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. used educational datamining to improve graduate students' performance, and overcome the problem of low grades of graduate students and try to extract useful knowledge from graduate students data collected from the college of Science and Technology. The data include fifteen years period [1993-2007]. After preprocessing the data, we applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, we present the extracted knowledge and describe its importance in educational domain.

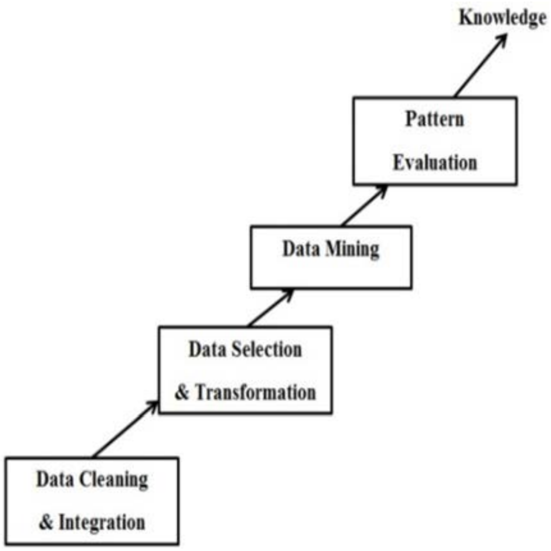
SonaliAgarwal, G. N. Pandey, and M. D. Tiwari (2012) [6] describes the educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation. Data Mining is anemerging technique with the help of this one can efficiently learn with historical data and use that knowledge for predicting future behavior of concern areas. Growth of current education system is surely enhanced if data mining has been adopted as a futuristic strategic management tool. The Data Mining tool is able to facilitate better resource utilization in terms of student performance, course development and finally the development of nation's education related standards. Monika Goyal and RajanVohra (2012) [7] applied data mining techniques to improve the efficiency of higher education institution. If data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution. This is an approach to examine the effect of using data mining techniques in higher education. Surjeet Kumar Yadav, BrijeshBharadwaj, and Saurabh Pal (2012) [11] used decision tree classifiers are studied and the experiments are conducted to find the best classifier for retention data to predict the student's drop-out possibility. Brijesh Kumar Baradwaj and Saurabh Pal (2011) [12] Used the classification task on student database to predict the students division on the basis of previous database. K.ShanmugaPriya and A.V.Senthil Kumar(2013) [13] applied a Classification Technique in Data Mining to improve the student's performance and help to achieve the goal by extracting the discovery of knowledge from the end semester mark. Bhise R.B, Thorat S.S and Supekar A.K. (2013) [14] used data mining process in a student's database using K-means clustering algorithm to predict students result. Varun Kumar and AnupamaChadha (2013) [15] used of one of the data mining technique called association rule mining in enhancing the quality of students' performances at Post Graduation level.Pallamreddy. venkatasubbareddy and VudaSreenivasarao (2010) [16] explained the Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal and use of decision trees is as a descriptive means for calculating conditional probabilities.

3. Data Mining Definition and Techniques

IJSER

Data mining refers to extracting or "mining" knowledge from large amounts of data [3]. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making [1]. The sequences of steps identified in extracting knowledge from data are: shown in Figure 1. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding.

Figure 1.The Steps of Extracting Knowledge from Data



3.1. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples [1]. In our case study we used ID3 decision tree to represent logical rules of student final grade.

3.2. Clustering

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. In educational data mining, clustering has been used to group students according to their behavior. According to clustering, clusters distinguish student's performance according to their behavior and activates. In this paper, students are clustered into three groups according totheir academics, punctuality, exams and soon [8]. 3.3. Association rule Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis [9].

3.4. Decision Trees

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal [10].

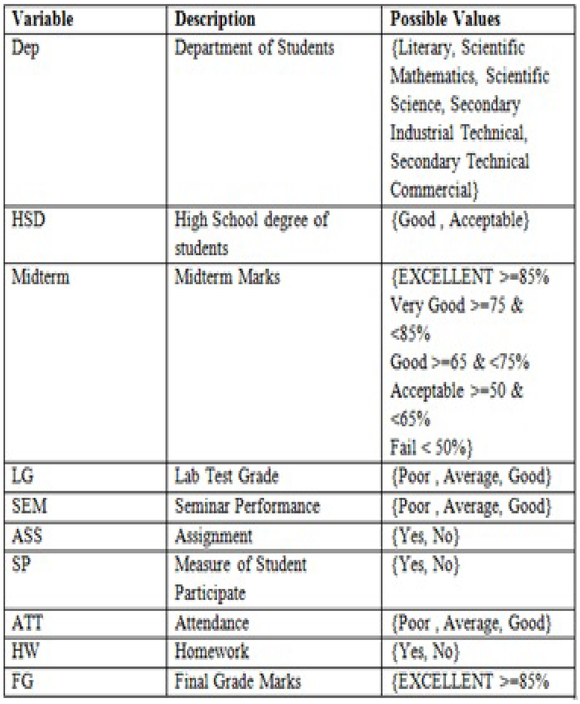
4. Data Mining Process
4.1. Data Preparations

The data set used in this study was obtained from a student's database used in one of the educational institutions, on the sampling method of Information system department from session 2005 to 2010. Initially size of the data is 1547 records. In this step data stored in different tables was joined in a single table after joining process errors were removed.

4.2. 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 Figure 2.

IJSER



Variable	Description	Possible Values
Dep	Department of Students	{Literary, Scientific Mathematics, Scientific Science, Secondary Industrial Technical, Secondary Technical Commercial}
HSD	High School degree of students	{Good , Acceptable}
Midterm	Midterm Marks	{EXCELLENT >=85% Very Good >=75 & <85% Good >=65 & <75% Acceptable >=50 & <65% Fail < 50%}
LG	Lab Test Grade	{Poor , Average, Good}
SEM	Seminar Performance	{Poor , Average, Good}
ASS	Assignment	{Yes, No}
SP	Measure of Student Participate	{Yes, No}
ATT	Attendance	{Poor , Average, Good}
HW	Homework	{Yes, No}
FG	Final Grade Marks	{EXCELLENT >=85%

Figure 2.Student Related Variables

IJSER

4.3. Decision Tree

A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. 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 distinct outcomes of the test. Each leaf node has a class label associated with it [11].

4.4. The ID3 Decision Tree

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric---information gain.
To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric - information gain. To find an optimal way to classify a learning set we need some function which provides the most balanced splitting. The information gain metric is such a function. Given a data table that contains attributes and class of the attributes, we can measure homogeneity of the table based on the classes. The index used to measure degree of impurity is Entropy [2]. The Entropy is calculated as follows: Splitting criteria used for splitting of nodes of the tree is Information gain. To determine the best attribute for a particular node in the tree we use the measure called Information Gain. The information gain, Gain (S, A) of an attribute A, relative to a collection of examples S, is defined as:

Entropy:

Entropy(S)=-P(positive)log2P(positive) -
P(negative)log2P(negative)

P(positive): proportion of positive examples in S

P(negative):proportion of negative examples in S

For example, if S is (0.5+, 0.5-) then Entropy(S) is 1, if S is (0.67+, 0.33-) then Entropy(S) is 0.92, if P is(1+, 0-) then Entropy(S) is 0. Note that the more uniform is the probability distribution, the greater is
its information.

Information Gain: measuring the expected reduction in Entropy
As we mentioned before, to minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice.

We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node.

The information gain, Gain(S,A) of an attribute A

Gain(S,A)= Entropy(S) -Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv)
We can use this notion of gain to rank attributes and to build decision trees where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root.

Advantages of using id3:

- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.
- Finding leaf nodes enables test data to be pruned, reducing number of tests.
- Whole dataset is searched to create tree.

Disadvantages of using id3:

- Data may be over-fitted or over-classified, if a small sample is tested.
- Only one attribute at a time is tested for making a decision.
- Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

4.5 How we implement ID3 Algorithm here:

ID3 (Learning Sets S, Attributes Sets A, Attributes values V)Return Decision Tree.

```
Begin
Load learning sets first, create decision tree root node 'root Node', add learning set S into root node as its subset.
For root Node, we compute
Entropy(rootNode.subset) first
If Entropy(root Node.subset)==0,
then rootNode.subset consists of records all with the same value for the categorical attribute, return a leaf node with decision attribute: attribute value;
If Entropy(rootNode.subset)!=0, then compute information gain for each attribute left(have not been used insplitting), find attribute A with Maximum(Gain(S,A)).
Create child nodes of this root Node and add to root Node in the decision tree.
For each child of the rootNode, apply ID3(S,A,V) recursively until reach node that has entropy=0 or reach leaf node.
End ID
```

This process goes on until all data classified perfectly or run out of attributes. The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules as shown in Table 2.TheTable 2discusses 8cases:

Case 1 - If Midterm Mark = Excellent, Lab Test Grade = Good, Student Participate = No, Homework = No, Seminar Performance = Good, Department = Scientific Mathematics then Final Grade = Very Good. Case 2 – If Midterm Marks = Excellent, Lab Test Grade = Good, Student Participate = No, Attendance = Good, Homework = No, Department = Secondary Technical Commercial then Final Grade = Very Good. Case 3 - If Midterm Marks = Excellent, Lab Test Grade = Good, Student Participate = No, Attendance = Good, Homework = No, Department = Secondary Industrial Technical then Final Grade = Very Good. Case 4 - If Midterm Mark = Excellent, Lab Test Grade = Poor, Attendance = Good then Final Grade = Very Good. Case 5 - If Midterm Mark = Excellent, Lab Test Grade = Average, Attendance = Good then Final Grade = Excellent. Case 6 - If Midterm

IJSER

Mark = Excellent, Lab Test Grade = Average, Attendance = Poor then Final Grade = Very Good. Case 7 - If Midterm Mark = Very Good, Lab Test Grade = Good, Homework = No, Seminar Performance = Good, Student Participate = No then Final Grade = Very Good.Case 8 - If Midterm Mark = Very Good, Lab Test Grade = Good, Homework = No, Seminar Performance = Good, Student Participate = No, Department = Scientific Mathematics then Final Grade = Very Good.

Example of ID3

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 2 for reference.

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

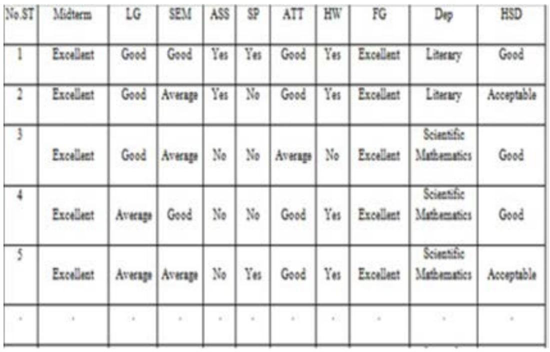
The symbolic attribute description

DEP={literacy, scientific mathematics,scientific science,secondary industrial technical, secondary technical commercial}

HSD={good, acceptable}
MIDTERM={EXCELLENT>=85%, VERY GOOD>=75&<85%, GOOD>=65&<75%,ACCEPTABLE>=50&<60%,FAIL<50%}
LG={poor,average,good}
SEM={poor,average,good}
ASS={yes, no}
SP={yes, no }
ATT={poor,average,good}
HW={yes, no}
FG={EXCELLENT>=85%}

5. Results and Discussion

The data set used in this study was obtained from a student's database used in one of the educational institutions, on the sampling method of Information system department from session 2005 to 2010. Initially size of the data is 1548 records are given in Figure 3.



No.ST	Midterm	LG	SEM	ASS	SP	ATT	HW	FG	Dep	HSD
1	Excellent	Good	Good	Yes	Yes	Good	Yes	Excellent	Literary	Good
2	Excellent	Good	Average	Yes	No	Good	Yes	Excellent	Literary	Acceptable
3	Excellent	Good	Average	No	No	Average	No	Excellent	Scientific Mathematics	Good
4	Excellent	Average	Good	No	No	Good	Yes	Excellent	Scientific Mathematics	Good
5	Excellent	Average	Average	No	Yes	Good	Yes	Excellent	Scientific Mathematics	Acceptable
.	.	.	.	.	.	.	.	.	.	.

Figure 3.Data Set

To work out the information gain for A relative to S, we first need to calculate the entropy of S. Here S is a set of 1547 examples are 292 " Excellent ", 536 "Very Good", 477 "Good", 188 "Acceptable" and 54 "Fail".

From the above rule set it was found that, Economic Status, Family and Relation Support, and other factors are of high potential variable that affect student's performance for obtaining good performance in examination result. The Confusion matrix is morecommonly named as Contingency Table. TABLE VII shows four classes and therefore it forms 4x4 confusion matrix, the number of correctly classified instances is the sum of diagonals in the matrix (292+536+477+188=1547); remaining all are incorrectly classified instances.

TABLE 1: CONFUSION MATRIX FOR MIDTERM USING ID3 ALGORITHM

result		Id3				
		excellent	Very good	good	acceptable	fail
Actual class	excellent	292	522	456	95	60
	Very good	500	536	268	159	55
	good	404	455	477	170	59
	acceptable	150	120	170	188	50
	fail	60	50	45	55	54

Table 2.Rule Set generated by Decision Tree

IJSER

IF Midterm='Excellent' AND LG='Good' AND SP='No' AND HW='No' AND SEM='Good' Dep='Scientific Mathematics' THEN FG='Very Good'
IF Midterm='Excellent' AND LG='Good' AND SP='No' AND ATT='Good' AND HW='No' AND Dep=' Secondary Technical Commercial' THEN FG='Very Good'
IF Midterm='Excellent' AND LG='Good' AND SP='No' AND ATT='Good' AND HW='No' AND Dep=' Secondary Industrial Technical' THEN FG='Very Good'
IF Midterm='Excellent' AND LG='Poor' AND ATT='Good' THEN FG='Very Good'
IF Midterm='Excellent' AND LG='Average' AND ATT='Good' THEN FG='Excellent'
IF Midterm='Excellent' AND LG='Average' AND ATT='Poor' THEN FG='Very Good'
IF Midterm='Very Good' LG='Good' AND HW='No' AND SEM='Good' AND SP='No' THEN FG='Very Good'
IF Midterm='Very Good' LG='Good' AND HW='No' AND SEM='Good' AND SP='No' AND Dep='Scientific Mathematics' THEN FG='Very Good'

IJSER

6. Conclusion

In this paper, decision tree method is used on student's database to predict the student's performance on the basis of student's database. We use some attribute were collected from the student's database to predict the final grade of student's. This study will help the student's to improve the student's performance, to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time.

Mrs D.BHU LAKSHMI ,Assistant professor is currently working in KNSIT College ,Department of CSE,Bangalore,India,PH-09379160729. E-mail: dasarilakshmimtech@gmail.com.
Mrs S.ARUNDATHI, Assistant professor is working in KNS IT College, Department of Computer Applications, Bangalore,India,PH-08897645007,E-mail:arundhathi28@gmail.com.
DR.JAGADEESH, Associate Professor, working in SRINIVASA INSTITUTE OF TECHNOLOGY, Amalapuram, Andhra Pradesh, India.

REFERENCES

[1] Brijesh Kumar Baradwaj, Saurabh Pal, Data mining: machine learning, statistics, and databases, 1996.

[2] Nikhil Rajadhyax, RudreshShirwaikar, Data Mining on Educational Domain, 2012.

[3] JiaweiHan, MichelineKamber, Data Mining: Concepts and Techniques, 2nd edition, 2006.

[4] Alaa El-Halees, Mining Students Data to Analyze Learning Behavior: A Case Study, 2008.

[5] Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students' Performance: A Case Study, 2012.

[6] SonaliAgarwal, G. N. Pandey, and M. D. Tiwari, Data Mining in Education: Data Classification and Decision Tree Approach, 2012.

[7] Monika Goyal ,Rajan Vohra2, Applications of Data Mining in Higher Education, 2012.

[8] P. Ajith, M.S.S.Sai, B. Tejaswi, Evaluation of Student Performance: An Outlier Detection Perspective, 2013.

[9] Varun Kumar, AnupamaChadha, An Empirical Study of the Applications of Data Mining Techniques in Higher Education, 2011.

[10] Hongjie Sun, Research on Student Learning Result System based on Data Mining, 2010.

[11] Surjeet Kumar Yadav, BrijeshBharadwaj, and Saurabh Pal, Mining Education Data to Predict Student's Retention: A comparative Study, 2012.

[12] Brijesh Kumar Baradwaj, Saurabh Pal, Mining Educational Data to Analyze Students‟Performance, 2011.

[13] K.ShanmugaPriya, A.V.Senthil Kumar,

[14]Improving the Student's Performance Using Educational Data Mining, 2013.

[15] Bhise R.B, Thorat S.S, Supekar A.K, Importance of Data Mining in Higher Education System, 2013.

[16] Varun Kumar, AnupamaChadha, Mining Association Rules in Student's Assessment Data, 2012.

[17] Pallamreddy.venkatasubbareddy, VudaSreenivasarao, The Result Oriented Process for Students Based On Distributed Data Mining, 2010