ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 4, June 2012

LEARNING BEHAVIOR OF ANALYSIS OF HIGHER STUDIES USING DATA MINING

Santosh kr. Gupta,
(Noida Institute of Engineering and Technology, Gr. Noida)
E Mail : **Skg31st@gmail.com**

Dr. Abhay Bnasal
Amity University Noida, India
(Head Dept. of Information Technology)
abansal1@amity.edu

Mr. Ritesh Rastogi
(Noida Institute of Engineering and Technology, Gr. Noida)
Associate Prof. (Computer Application)

ABSTRACT

The main concern of providing higher education is to provide quality education to the students and to produce technically qualified professionals. The knowledge is hidden among the educational data set and it is extractable through data mining techniques. Educational data mining methods from discovering knowledge from data that we used from educational data mining to analyze learning behavior. We have taken data of post Graduate students (MCA) and applied data mining techniques.

Keywords: Educational Data Mining (EDM), Learning Management System

1. INTRODUCTION

Data mining is an important paradigm for educational assessment. The usual assumption is that mining is performed after educational activity with that activity having been designed without regard for the mining process [1]. Educational data mining (EDM) is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context [2]. Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data mining is used in most areas where data are collected-marketing, health, communications, education etc. Data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. Data mining techniques have been introduced into new fields of Statistics, Databases, Machine Learning, Pattern reorganization, Artificial Intelligence and computation capabilities etc [3]. There are many data mining techniques, most of the work that has been done in higher education falls into the Categories of clustering, classification, visualization, and association analysis. Using these techniques many kinds of knowledge can be discovered The discovered knowledge can be used for prediction regarding Enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students performance and so on [3].

2. DATA MINING DEFINITION AND TECHNIQUES

Data mining and knowledge discovery in database is a collection of exploration techniques based on advanced analytical methods or tools for handling a large amount of information . The techniques can found novel patterns that may assist an enterprise in understanding the business better and in forecasting.

Data mining is the collection of techniques for efficient automated discovery of previously unknown, valid, novel, and useful and understandable patterns in large databases. The pattern must be actionable so that may be used in an enterprise's decision making process.[13] The sequences of steps

identified in extracting knowledge from data are shown in Figure 1.

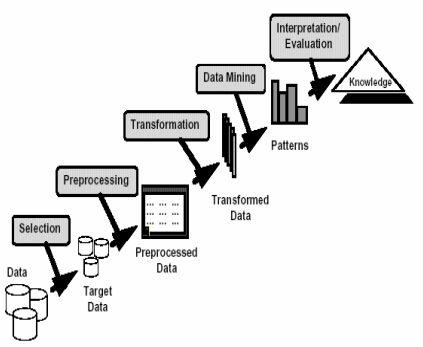

Figure 1: The steps of extracting Knowledge from data

Data mining employs number of techniques like Classification, Clustering, and Association rules, Decision trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

2.1 Classification

Classification is separation or ordering of objects (or things) into classes. If classes are created without looking at the data, the classification is called apriori classification. If classes are created with looking at data then classification is called posteriori classification.

2.2 Clustering

Clustering is "the process of organizing objects into groups whose members are similar in some way".
Clustering can be considered the most important unsupervised learning technique; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. Clustering is used for simplification, pattern detection, data concept constructions, unsupervised learning process.

2.3 Association rules

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

2.4 Decision trees

Decision tree is a popular classification method that results in flow-chart like tree structure where each node denotes a test on an attribute value and each branch represents an outcome of the test. It displays relationships found in training data. The consists of zero or more internal nodes and one or more leaf nodes with each internal

3. Related work

J.F.Superby and J.P.Vandamme determine the factors that influence the first year students results using the data mining methods [8].
Brijesh Kumar Baradwaj and saurabh Pal analyze the students data and predict their end semester division[3].Cesar Vialardi provide a solution of a problem that is faced by university students is to take the right decision in relation to their academic itinerary based on available information . He proposes the use of a recommendation system based on data mining technique to help students to take the decisions on their academic itineraries [9].
Romero C., Ventura, using the data mining for course management system [10].
Merceron, A. and Yacef K. Used data mining to warn failing students before final exam [11]. Pandey and pal using the Bayes classification and analyze the student data and found that whether new comer students will performer or not.[12]

4.Data mining Methodology

Many studies were undertaken in order to try to explain the academic performance or to predict the success or the failure. But we have seen that Students performance can be predict On the bases of secondary, higher secondary, graduate, class test, medium of studies, teacher assessment and their class attendance.

4.1 Data Evaluation process

We have collected data of MCA students for the course held by MTU university for first semester students. The number of students was 60. The data we collected included Name, their enrollment number along with their personal and previous educational background. We use registration form for getting the information.

4.2 Data preparation

After collecting the data we synchronized it and converted into excel format. We also took their attendance, their class test marks, Teacher assessment. Students training data and related variable are given in Table 1.

s no	variables	Possible values
1	X	{first div., second div.}
2	Xii	{first div., second div.}
3	Graduate	{First div. second div.}
4	Medium	{Hindi, English}
5	stream	{BCA, B.Sc.}
6	Teacher Assessment (TA)	{Excellent, Good, Avg, Bad}
7	Attendance (ATT)	{Good , Avg, Poor}
8	Class Test(CT)	{Excellent, Good, V.good, Avg, Poor}
9	End Sem. Grade(ESG)	{Excellent, Good, Avg, Bad}

Table 1

Target value definitions:

- **X-** It is split into two class values : first -> 60% , second>45% and <60%
- **XII -** It is split into two class values : first -> 60% , second>45% and <60%
- **Graduate--** It is split into two class values : first -> 60% , second>45% and <60%
- **Teacher Assessments-** it is depends on the presentation skill, behavior of student, assignments. It is divided into four categories Excellent, good, average, bad.

- **Attendance-** it is split into Three class variables: Good>=60,Avg<60 and >=45,Poor<45**.**

- **Class Test** It is split into five classes values :
  Excellent ->=75% , V.Good<75% and >=65%, Good<65% and >=60% Avg<60 and >=45%, Poor<45%

- **End Sem. Marks** It is split into four class values :
  Excellent ->=75% , Good<75% and >=60%,Avg<60 and >=45%, Bad<45%

5. Building A Decision Tree- The tree Induction algorithm

Decision Tree is a model that is both predictive and descriptive. It displays relationships found in training data. The consists of zero or more internal nodes and one or more leaf nodes with each internal. The training process that generates the tree is called induction. Complexity of decision tree increases as the number of attributes increases; although in some situations it has found that only a small number of attributes can determine the class to which an object belongs and rest of attributes have little

impact. Decision tree algorithm is top-down greedy algorithm. The aim of the algorithm is to build a tree that has leaves that are as homogenous as possible.
The decision tree algorithm is given below:

1. Let the set of training data be S. if some of the attributes are continuous valued, they are discretized.

2. If all instances in S are in same class, then stop.

3. Split the next node by selecting an attribute A from amongst the independent attributes that best divides or splits the objects in the node into subsets and create a decision tree node.

4. Split the node according to values of A.
5. Stop if either of following conditions is met , otherwise continue with step 3:
   (a) If this partition divides the data into subsets that belong to a single class no other node needs splitting.
   (b) If there is no remaining attributes on which the sample may be further divide.[13]

To find a split attribute, two most Commonly used evaluation rules:

- Rules based on information theory
- Rules based on Gini index.

5.1 Split algorithm based on information Theory

Information is defined as $-p_i \log p_i$ where p_i is the probability of some event. Since p_i is always less than 1.
$\log p_i$ is always negative and $-p_i \log p_i$ is always positive. Information of any event that is likely to have severable possible outcomes is given by

$$I=\sum_i (-p_i \log p_i)$$

5.1.1 Information Gain

It is a measure of how good an attribute is for predicting the class of each of training data. We will select highest information gain as the next split attribute. Assume there are two classes , p and N, and let the training data S(with the total number of objects s) contain p elements of class p and n element of class N. The amount of information is defined as

$$I= -(n/s) \log(n/s) - (n/s) \log(n/s)$$

We define information gain for sample S using attribute A as follows:

$$Gain(S, A) = I - \sum_{i \in values(A)} (t_i/s) I_i$$

I is the information before split and
$\sum_{I \in values(A)} (t_i/s) I_i$ sum of information after the split where I_i is the information of node i and t_i is the number of objects in node i

6. Results

We have collected data of MCA students for the course held by MTU university for first semester students. Applying the decision tree algorithms and knowledge represented in the form of IF-THEN rules. Some of strong rules in the tree are:

IF CT='Excellent' And TA=' Excellent' And stream ='BCA' THEN ESG=' 'Excellent'
IF CT='Excellent' And TA='Good' And X='Second' THEN ESG='Good'
IF CT='V.Good' And Xii= second And stream =' B.Sc ' Or stream =' BCA ' THEN ESG='Avg'
IF CT='Good' And ATT='Poor' ' THEN ESG='Avg'
IF CT=' Avg' And TA='Avg' THEN ESG='Bad'
IF CT=' Poor' And Graduate ='First' And TA='Bad' And Medium='English' THEN ESG=' Avg'

7. Conclusion

In this paper we have try to give some rules that can help to predict the students marks

quality on the basis of their previous records. For predicting the marks we have used decision tree method of classification. Information's like class attendance, teacher assessment, secondary marks, higher secondary marks, graduate marks, stream of graduate, medium of study were collected for predicting the end semester marks quality.

This study will help both students and teachers to improve the quality of result of end semester.

7. References:

[1] Steven L. Tanimoto, University of Washington Dept. of Computer Science and Engineering "Improving the Prospects for Educational Data Mining"

[2] Crist´obal Romero, and Sebasti´an Ventura 'Educational Data Mining: A Review of the State of the Art' IEEE TRANSACTIONS ON SYSTEMS, MAN, AND YBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 40, NO. 6, NOVEMBER 2010

[3] Brijesh Kumar Baradwaj and Saurabh Pal "Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[4] RYAN S.J.D. BAKER Worcester Polytechnic Institute Worcester, MA USA and KALINA YACEF School of Information Technologies ' The State of Educational Data Mining in 2009: A Review and Future Visions'

[5] Saleema Amershi and Cristina Conati 'Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments' JEDM Volume 1, Issue 1, Pages 18-71

[6] Tara Madhyastha and Earl Hunt 'Mining Diagnostic Assessment Data for Concept Similarity' JEDM Volume 1, Issue 1, Pages 72-91

[7] D'Mello, S. and Olney, A. and Person, N. 'Mining Collaborative Patterns in Tutorial Dialogues' D'Mello, S. and Olney, A. and Person, N., JEDM Volume 2, Issue 1, Pages 1-37

[8] J.F. Superby, J-P. Vandamme, N. Meskens. "Determination of factors influencing the achievement of the first-year university students using data mining methods". Workshop on Educational Data Mining 2006.

[9] Cesar Vialardi, Javier Bravo, Leila Shafti, Alvaro Ortigosa (2009) Recommendation in Higher Education Using Data Mining Techniques In: Proceedings of Educational Data Mining 2009 Edited by:Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (Eds.). 190-199 International Working Group on Educational Data Mining

[10] Romero,Cristobal; Ventura, Sebastian; Garcia, Enrique, Computers & Education, vol. 51 ,No.1. pp. 368-384 Aug 2008

[11] Merceron A. and Yacef K 'Educational Data Mining: a Case Study' In Proceedings of 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press 2005.

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

[13] G.K.Gupta 'Introduction to Data Mining with Case Studies" PHI Learning Private Limited