RESEARCH ARTICLE OPEN ACCESS

Use of Data Mining to Analyse Students' Performance

Shubham Garg[1], Trapti Gupta[2], Shubhankar Gupta[3], Mr. Prashant S. Chavan[4]
[1,2,34](Information Technology, BharatiVidyapeeth Deemed to be University College of Engineering, Pune)

Abstract:

In today's world of computerization, education is not limited to old lecture method. Highest level of quality in higher education can be achieved by discovering knowledge for students' performance prediction. This knowledge can be extracted using data mining techniques. This paper is to justify how data mining techniques can be used in context of higher education. In this research the decision tree method is used for data classification and to evaluate students' performance. Using decision tree method we can describe students' performance in end semester examination.

Keywords — **Data Mining (EDM); Classification; Data Mining process; Decision Tree.**

I. INTRODUCTION

The large volumes of data storage in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats are used to approach of information technology in various fields. 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]. In data mining we apply various methods and algorithms to discover and extract patterns of stored data [2]. Data mining helps significantly in decision making due to which it has become essential component in various organizations.

Analysing and understanding the factors for poor performance is a complex and incessant process hidden in past and present information congregated from academic performance and students' behaviour. Powerful tools are required to analyse and predict the performance of students scientifically. If, Universities could identify the factors for low performance earlier and is able to predict students' behaviour, this knowledge can help them in taking pro-active actions. Students will be able to identify their weaknesses beforehand and can improve themselves. Teachers will be able to plan their lectures as per the need of students and can provide better guidance to such students. Parents will be reassured of their ward performance in such institutes.

Analysis and prediction with the help of data mining techniques have shown noteworthy results in the area of fraud detection, predicting customer behavior, financial market, loan assessment, bankruptcy prediction, real-estate assessment and intrusion detection. It can be very effective in Education System as well.

There is rapid increasing interest in using data mining in education. Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments [3]. Educational Data Mining uses many techniques such as Naïve Bayes, Neural Networks, Decision Trees, K- Nearest neighbor, and many others. Many kinds of knowledge can be discovered by using these techniques such as association rules, classifications and clustering. The discovered knowledge can be used for prediction about students' performance.

We will use data mining techniques to study students' performance in the courses. Data mining can efficiently be used to study the performance of

the students. The classification task of data mining is used to evaluate students' performance. In this research, we have particularly used decision tree method to predict whether the students will pass according to their end semester marks. We have taken information like students' Attendance, Class Test Grade and Assignment marks from the students' database from the university. This information is analyzed and as a result a decision tree is made for predicting students' performance.

II. DATA MINING DEFINITIONS AND TECHNIQUES

Data mining is also known Knowledge Discovery in Databases. It refers to extracting or "mining" knowledge from large databases. These databases contains large amount of data. Data mining techniques are used to discover relationships and hidden patterns. These hidden patterns help in decision making. Many times knowledge discovery and data mining are treated as synonyms but data mining is actually a part of knowledge discovery. The series of steps identified in extracting knowledge from data are shown in Figure 1.

Classification, Regression, Clustering, Neural Networks, Association Rules, Artificial Intelligence, Decision Trees, and Genetic Algorithm, Nearest Neighbor method etc. are various algorithms and techniques which are used for knowledge discovery from databases. We need to have brief understanding of these techniques and methods.

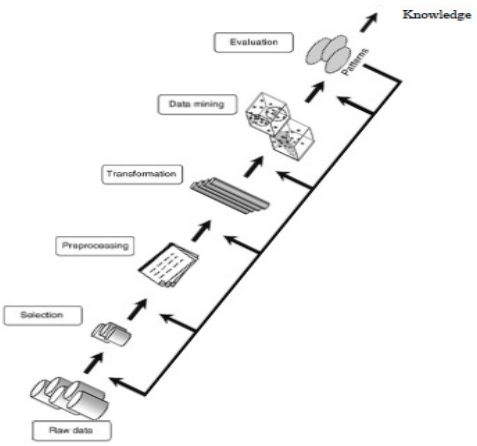


Figure 1: The steps of extracting knowledge from data

A. *Classification*

The most commonly applied data mining technique is Classification. Classification gives us a set of pre-classified examples which develop a model. This model can classify record of population at large scale. This approach frequently involves decision tree or neural network-based classification algorithms. The process of data classification involves learning and classification. Classification algorithm is used to analyze training data in learning. For example, a bank loan officer wants to analyze data in order to know which loan applicants are risky or which are safe.

B. *Clustering*

Clustering can identify the object of the same classes. This technique is used to identify dense and sparse region in object space. It is also used to discover data distribution pattern and it correlates the attributes of data. In classification, it becomes costly to distinguish between groups or classes of object, that's why we use clustering as a preprocessing approach for classification and attribute subset selection.

C. *Prediction*

For predicting the data, we can use regression technique. By Regression analysis we get to know the relation between one or more independent variables and dependent variables. There are independent attributes that are already known and response variables are the one which we have to predict in data mining. Prediction cannot solve many real-world problems. Forecasting future values, it is necessary to have more complex techniques (e.g. logistic regression, decision trees, or neural nets). For both regression and classification the same model can be used.

D. *Association Rule*

Association and correlation are used to find item set available in large data sets. Catalogue design, cross marketing and customer shopping behavior analysis etc. are the type of finding helps in businesses to make certain decisions. Association rule algorithms

generate a rule with confidence values less than one. Specific dataset is having the number of possible Association Rules which are generally very large and a high proportion of the rules having usually of little (if any) value.

E. *Neural Networks*

Neural network is a set of connected input/output units and each and every connection has some weight present with it. During the learning phase, adjusting weights is a part of network learns so that it is able to predict the correct class labels of the input tuples. Remarkable ability to derive meaning from complicated or imprecise data is in the list of neural networks. This can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

F. *Decision Tree*

Decision tree is tree-shaped structures which is basically representing sets of decisions. Rules for the classification of a dataset are generated by decisions. Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) are some of Specific decision tree methods.

G. *Nearest Neighbor Method*

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k >= 1). It is also called as the k-nearest neighbor technique.

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 teacher gives students internal assessment marks based on their performance in activities like attendance, class tests and assignments. The end semester examination is one that is scored by the student in semester examination. If student scores more than minimum marks required passing a semester in internal assessment then that student goes to next semester.

A. *Data Preparations*

The data set used in this study was obtained from BharatiVidyapeeth University, College of Engineering, Pune on the sampling method of Information Technology department of course B.Tech (Bachelors of Technology) from session 2014 to 2018. Initially size of the data is 20.

B. *Data Selection and Transformation*

Fields required for data mining are selected in this step. 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 I for reference. Variables used for the investigation are as follows:

CTG **–** Class test grade obtained. Two class tests are conducted each semester and average of these tests is calculated. CTG is split into three classes: *Poor – < 40%, Average – > 40% and < 60%, Good –>60%.*

ASS **–** Assignment performance. In each semester six assignments for each subject are given to students. Assignment performance is divided into two classes: *Yes – student submitted assignment, No – Student not submitted assignment.*

Variable	Description	Possible
CTG	Class Test Grade	{Poor, Average, Good}
ASS	Assignment	{Yes, No}
ATT	Attendance	{Poor, Average, Good}
ESM	End Semester Mars	{Pass, Fail}

ATT **–** Attendance of Student. Minimum 75% attendance is compulsory to participate in End Semester Examination. But even through in special cases low attendance students also participate in End Semester Examination on genuine reason. Attendance is divided into three classes: *Poor - <60%, Average - > 60% and <80%, Good - >80%.*

ESM **-** End semester Marks. Marks obtained in B.Tech.semester and it is declared as response variable. It is split into two class values: Pass– *>45% and Fail < 40%.*

C. Decision Tree

A decision tree contains a root node, branch node and leaf node. Branch node represents choices between numbers of alternatives. Leaf node represents a decision.

Decision tree are generally used for the purpose of decision-making. Decision tree starts with a root node. It is on user to take actions. From root node, a user extends the tree recursively according to the algorithm. This finally gives us a decision tree in which every branch represents a scenario of decision and its outcome. We will be using ID3 decision tree learning algorithm.

D. The ID3 Decision Tree

Construction of the decision tree is the basic idea of ID3 algorithm by using a top-down, greedy search for checking the given sets to test the each attribute at every tree node. Selection of the attribute in specific order is most useful for classifying a given sets, for this metric - information gain is introduced.

Classification of a learning setwe have to find an optimal way.The best way is to minimizing the depth of the tree. For measuring, we need some function which provides the most balanced splitting. E.g. information gain metric.

E. Measuring Impurity

In the above data table which contains attributes and class of the attributes, which can be used to measure heterogeneity or homogeneity of the table based on the classes. Table is said to be pure or homogenous if it contains only a single class and if table contains several classes, then the table is called impure or heterogeneous. The degree of impurity can be measured by several indices.Gini index, entropy, and classification errorare most well-known indices to measure degree of impurity.

$$\text{Entropy} = \sum_j - p_j \log_2 p_j$$

If probability is 1 and log (1) = 0 thenthe entropy of a pure table (consist of single class) is zero.
When all classes in the table have equal probability then entropy reaches to the maximum value.

F. Splitting Criteria

After a dataset is split on an attribute the information gain depends on the decrease in entropy. In the construction of a decision tree, it is mainly about finding attribute which returns the highest information gain. The information gain, relative to a collection of examples S,Gain (S, A) of an attribute A, is defined as

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)|$$

Where,
Values (A) = set of all possible values for attribute *A* .*Sv* = subset of *S* for which attribute *A* has value *v*.

Gain is the entropy of the original collection *S in* the first term in the equationand the second term is the expected value of the entropy after S is partitioned using attribute *A*. The second term is simply the sum of the entropies of each subset, weighted by the fraction of examples $|S_v|/|S|$ that belong to *Gain (S, A)* is therefore the expected reduction in entropy is changed by knowing the value of attribute *A*.

This process continues for each new leaf until either of two conditions is met:

1. All the attribute are included along the path of the tree.
2. The examples are bound with the leaf node all have their entropy is zero.

IV. RESULTS AND DISCUSSION

The data set of 20 students used in this study was obtained from BharatiVidyapeeth College of Engineering, Pune (Maharashtra) Information Technology Department of course B.Tech.

S.No.	CTG	ASS	ATT	ESM
1.	GOOD	YES	GOOD	PASS
2.	GOOD	YES	GOOD	PASS
3.	GOOD	YES	AVG	PASS
4.	AVG	YES	POOR	PASS
5.	AVG	NO	GOOD	PASS
6.	POOR	NO	AVG	FAIL
7.	GOOD	NO	AVG	PASS
8.	AVG	NO	GOOD	PASS
9.	POOR	YES	GOOD	PASS
10.	POOR	YES	POOR	PASS
11.	GOOD	NO	AVG	PASS
12.	GOOD	YES	GOOD	PASS
13.	AVG	YES	GOOD	PASS
14.	AVG	NO	POOR	FAIL
15.	POOR	YES	GOOD	PASS
16.	GOOD	NO	POOR	FAIL
17.	AVG	YES	POOR	PASS
18.	AVG	NO	AVG	FAIL
19.	GOOD	YES	AVG	PASS
20.	POOR	NO	POOR	FAIL

By using the data given in the above table first we will find out the Information gain for the table.
Here, total passed students (p) = 14 and total failed student (n) = 6.

$IG(p,n) = -(p/p+n)\log_2(p/p+n) - (n/p+n)\log_2(n/p+n)$
$= -14/20(\log_2(14/20)) – 6/20(\log_2 6/20))$
$= (-0.7*-0.5146) + (-0.3*-1.7369)$
$IG(p,n) = 0.88129$

Now we will find Entropy of CTG, ASS and ATT.

CTG	PASS	FAIL	IG(P_i, F_i)
GOOD	7	1	**0.543**
AVG	5	2	**0.861**
POOR	1	4	**0.721**

$IG(7,1) = -7/8(\log_2(7/8)) – 1/8(\log_2(1/8)$
$= -0.875*\log_2 0.875 – 0.125*\log_2 0.125$
$= 0.5435$

$IG(5,2) = -5/7(\log_2(5/7)) – 2/7(\log_2(2/7)$
$= -0.714*\log_2 0.714 – 0.2857*\log_2 0.2857$
$= 0.861$

$IG(1,4) = -1/5(\log_2(1/5)) – 4/5(\log_2(4/5)$
$= -0.2*\log_2 0.2 – 0.8*\log_2 0.8$
$= 0.721$

Now, We will find Entropy for CTG.

$E(CTG) = \sum (p_i + n_i/p+n)(IG(p_i/n_i))$
$= 8/20\ (0.543) + 7/20\ (0.861) + 5/20\ (0.721)$
$= 0.2172 + 0.3013 + 0.1802$
$= 0.6987$

$Gain(CTG) = IG(p,n) – Entropy(CTG)$
$= 0.8813 – 0.6987$
$= 0.1826$

Similarly for ASS and ATT,

ATT	PASS	FAIL	IG(P_i, F_i)
GOOD	8	0	**0**
AVG	4	2	**0.917**
POOR	2	4	**0.917**

E(ASS) = 0.6871

Gain(ASS) = 0.1942

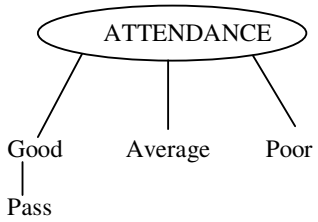
ASS	PASS	FAIL	IG(P_i, F_i)
YES	10	1	**0.439**
NO	4	5	**0.9906**

E(ATT) = 0.5502

Gain(ATT) = 0.3311

Gain of ATT is more than Gain of CTG and ASS. So Main Root Node will be Attendance node. As the $IG_{GOOD} = 0$. So, Good will always give us PASS.

Now we will find Gain of ASS, CTG with respect to ATT_{AVG}.



ATT	ASS	CTG	ESM
AVG	Yes	Good	**Pass**
AVG	No	Poor	**Fail**
AVG	No	Good	**Pass**
AVG	No	Good	**Pass**
AVG	No	Avg	**Fail**
AVG	No	Good	**Pass**

Entropy of CTG.

CTG	PASS	FAIL	IG(P_i, F_i)
GOOD	4	0	**0**
AVG	0	1	**0**
POOR	0	1	**0**

E(CTG) = 0

Gain(ATT_{AVG}, CTG) = 0.917

Entropy of ASS,

ASS	PASS	FAIL	IG(P_i, F_i)
YES	2	0	**0**
NO	2	2	**0.528**

E(ASS) = 0.667

Gain(ATT_{AVG}, ASS) = 0.917 – 0.667
= 0.25

Gain of ASS is more than Gain of CTG with respect to ATT_{AVG} the new node will be CTG under ATT_{AVG}. Now, under CTG we will refer the table for GOOD, AVG and POOR because it is clear when student will fail and when student will pass just by looking at the table.

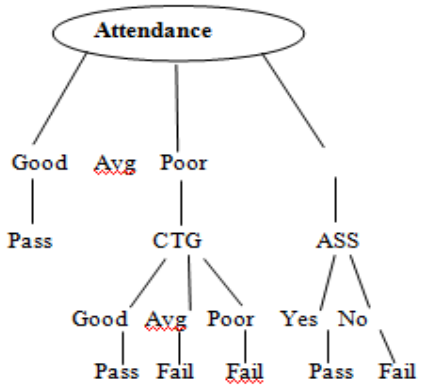
Now, we will find out Entropy and Gain of ASS with respect to ATT_{POOR}. We will not consider CTG as it is already used.

ASS	PASS	FAIL	IG(P_i, F_i)
YES	3	1	**0.5**
NO	0	3	**0**

E(ASS) = 0.286

Gain(ATT_{POOR},ASS) = 0.631

Next node will be of ASS under ATT_{POOR} Node. It will be the last node and we will decide PASS and FAIL by referring the table as it is clear just by seeing table.



The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules.

IF ATT = "GOOD" AND CTG = "GOOD" OR CTG = "AVG" THEN ESM = "PASS"
IF ATT = "AVG" AND CTG = "GOOD" THEN ESM = "PASS"
IF ATT = "AVG" AND CTG = "AVG" OR CTG = "POOR" THEN ESM = "FAIL"
IF ATT = "POOR" AND ASS = "YES" THEN ESM = "PASS"
IF ATT = "POOR" AND ASS = "NO" THEN ESM = "FAIL"

Rules Generated by Decision Tree

Classification Rules can be generated using the decision tree and these Classification rules defines the path from root node to terminal node. IF-THEN rules will describe when a particular student will pass and when he/she will fail. We have used Decision Tree to predict the student end semester result based on few internal assessment marks.

V. CONCLUSION

In this paper, the classification task is used on student database collected from BharatiVidyapeeth College of Engineering to predict the students division on the basis of collected database. The decision tree method is used here. Information like Class Test Grade (CTG), Assignment (ASS), Attendance (ATT) to predict the end semester result (ESM).

This study will help students and teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination.

VI. ACKNOWLEDGEMENT

We would like to acknowledge Saurabh Pal and other contributors for developing and maintaining the IJACSA'sMining Educational Data to Analyze Students' PerformanceJournal paper which have been used in the preparation of this paper.

VII. REFERENCES

1. Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996.
2. U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996.
3. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.
4. J. R. Quinlan, "Introduction of decision tree: Machine learn", 1: pp. 86106, 1986
5. Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.
6. Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. IJACSA - International Journal of Advanced Computer Science and Applications, 2(3), 80-84. Retrieved from http://ijacsa.thesai.org.