Predicting Students Attrition using Data Mining

Rakesh Kumar Arora
Department of Computer Science,
Krishna Engineering College,
Ghaziabad, UP, India

Dr. Dharmendra Badal
Department of Mathematical Science & Computer Applications,
Bundelkhand University,
Jhansi, UP, India

Abstract- Student attrition has become one of the most important measures of success for higher education institutions. It is an important issue for all institutions due to the potential negative impact on the image of the university and the institution and is great hindrance on the career path of the dropouts. A system to identify students that have high risk of attrition using Decision tree is being described in this paper. The paper also focuses on reasons on attrition of students and steps need to be taken to improve student's retention. The result of analysis will assist the institutions in predicting the set of students who can leave the institution after confirming admission and steps that need to be taken to improve student's retention.

Keywords: Student Attrition, Data Mining, Decision Tree, Information Gain, Entropy

I. INTRODUCTION

II.

During last two decades, number of Higher Education Institutions has grown rapidly in India leading to cut throat competition among these institutions. Most of the institutions are opened in self finance mode, so all time they feel short hand in expenditure. Therefore, these institutions focused more on the strength of students rather than on the quality of education.

Increasing student retention is a long term goal in all academic institutions. The consequences of student attrition are significant for students, academic and administrative staff. The most vulnerable students at all institutions of higher education are the first-year students, who are at greatest risk of dropping out in the first term or trimester of study or not completing their programme/degree on time. Therefore most retention studies address the retention of first-year students. Consequently, the early identification of vulnerable students who are prone to drop their courses is crucial for the success of any retention strategy. This would allow educational institutions to undertake timely and pro-active measures. Once identified, these 'at-risk' students can be then targeted with academic and administrative support to increase their chance of staying on the course. [1]

A very promising tool to attain valuable information about student's attrition is the use of data mining. Data mining techniques are used to discover hidden information, patterns and relationships of large amount of data, which is very much helpful in decision making. With the help of data mining methods, such as decision tree it is possible to identify the set of students that are likely to leave the institution.[2]

III. REASONS FOR ATTRITION OF STUDENTS

Attrition rates for classes taught through distance education are much higher than classes taught in a face-to-face setting. The reasons for attrition through distance education or face to face sitting may vary but some of the most prevalent reasons for leaving institutions include Dissatisfaction with academic matters (poor results in internal exams, not being able to cope up with stress that the course requires, change in career goals or inability to take desired course), financial difficulties, motivational problems, personal considerations (marriage, pregnancy, family responsibilities or illness),dissatisfaction with the institution and preference of full time jobs over study.[3]

IV. CONSEQUENCES OF ATTRITION OF STUDENTS

Student's attrition is a major concern for any type of educational institution as huge costs are incurred with respect to time, resources and tuition for students, faculty, institutions and other members of society. Whenever students drop out of programs, number of times their "seats" remain empty for the duration of the program resulting in loss of tuition fees for the institution and importantly when the program is complete, there will be fewer professionals entering the workforce. This is the area of concern especially in the medical field where there is the dearth of available and qualified professionals to meet the health care needs of society. [4]

V. METHODOLOGY

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node has two or more branches while leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data. The core algorithm for building decision trees called **ID3** by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy* and *Information Gain* to construct a decision tree. [5]

To find an optimal way to classify a learning set, one has 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.

Let node *N* represents the tuples of partition *D*. The attribute with the highest information gain is chosen as the splitting attribute for node *N*. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. The expected information needed to classify a tuple in *D* is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

-equation 1

where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. Info(D) is also known as the entropy of D.[6]

Suppose we were to partition the tuples in *D* on some attribute *A* having *v* distinct values, ($a1$, $a2$, … , a_v), as observed from the training data. Attribute *A* can be used to split *D* into *v* partitions or subsets, ($D1$, $D2$, …, $D_{v)}$,where D_j contains those tuples in *D* that have outcome a_j of *A*. These partitions would correspond to the branches grown from node *N*. One would like each partition to be pure but it is possible that the partitions may be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). Hence more information would be needed (after the partitioning) in order to arrive at an exact classification. This amount is measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

-equation 2

The term $|D_i|/|D|$ acts as the weight of the j^{th} partition. $Info_A(D)$ is the expected information required to classify a tuple from *D* based on the partitioning by *A*. The smaller the expected information required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on *A*). That is,

$$Gain(A) = Info(D) - Info_A(D).$$

-equation 3

Gain(A) tells us how much would be gained by branching on *A*. It is the expected reduction in the information requirement caused by knowing the value of *A*. The attribute *A* with the highest information gain, ($Gain(A)$), is chosen as the splitting attribute at node *N*.[6]

VI. ANALYSIS AND RESULTS

The study was carried out on the students that have taken admission in B.Tech in the year 2012 in reputed Engineering College of Ghaziabad. The attributes considered for analysis includes Division of students in qualifying exam, Family income, Sex, Admission through counseling or direct admission and whether the student have withdrawn admission or not. The division of students in qualifying exams have been classified in three categories: Distinction for students securing more than 75% marks, I division for students securing between 60% and 74.9% marks and II division for students securing less than 59.9% in qualifying exam. The attribute income has also being classified in three categories: High for the students having family income greater than 8 lakhs per annum, Medium for the students having family income between 5 lakhs per annum and 8 lakhs per annum and Low for the students having family income less than 5 lakhs per annum. The sample student database is shown in Table 1

TABLE 1: SAMPLE OF STUDENTS DATABASE

S. No.	Division	Income	Sex	Direct/ Counseling	Admission Withdrawn
1	II	High	M	Direct	No
2	II	High	M	Counseling	No
3	I	High	M	Direct	Yes
4	Distinction	Medium	M	Direct	Yes
5	Distinction	Low	F	Direct	Yes
6	Distinction	Low	F	Counseling	No
7	I	Low	F	Counseling	Yes
8	II	Medium	M	Direct	No
9	II	Low	F	Direct	Yes
10	Distinction	Medium	F	Direct	Yes
11	II	Medium	F	Counseling	Yes
12	I	Medium	M	Counseling	Yes
13	I	High	F	Direct	Yes
14	Distinction	Medium	M	Counseling	No

The expected information needed to classify a tuple in data partition (D) is determined by using equation 1

Info(D)= .940 bits

The expected information needed to classify the tuple in D if tuples are partitioned according to division is determined by using equation 2

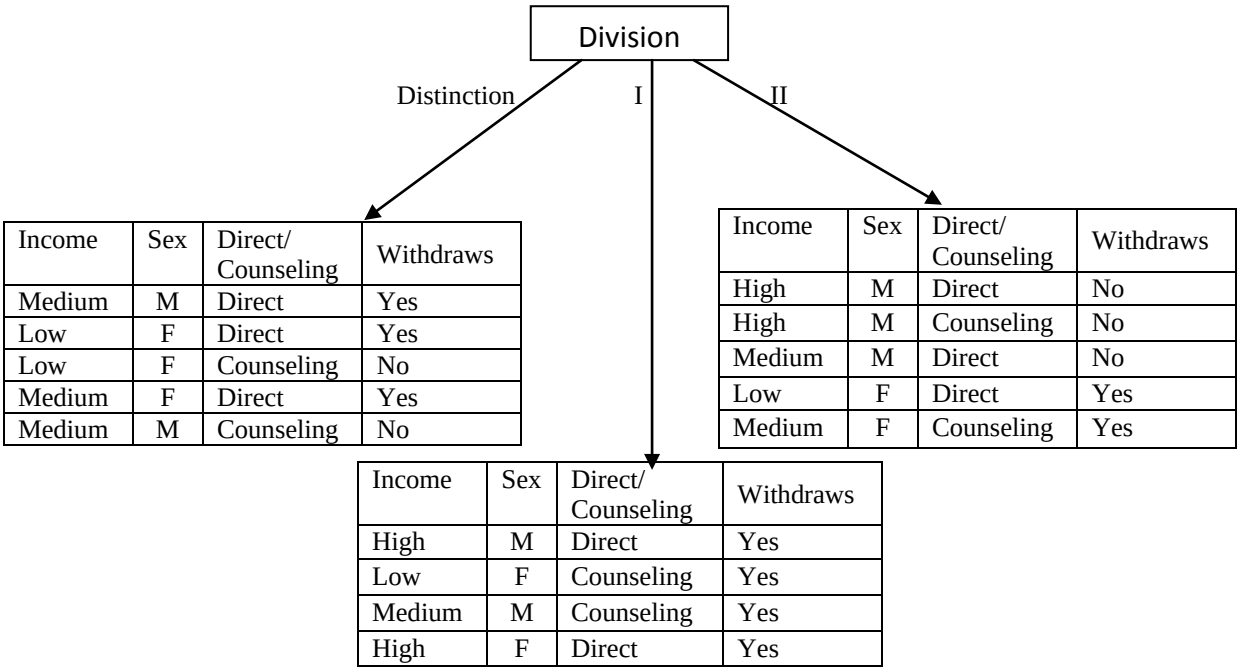
$Info_{income}(D)$=.694 bits

Therefore the gain in information from such partitioning would be

Gain (income)= Info(D) – $Info_{income}(D)$=.246 bits

Similarly Gain (income)=.029 bits, Gain(Sex)=.151 bits and Gain(Direct/Counseling) = .048 bits. Since division has the highest information gain among all attributes, it is selected as splitting attribute. The tuples are than partitioned accordingly as shown in Figure 1.

FIGURE 1: THE ATTRIBUTE DIVISION BECOMES THE SPLITTING ATTRIBUTE AT THE ROOT NODE OF DECISION TREE.

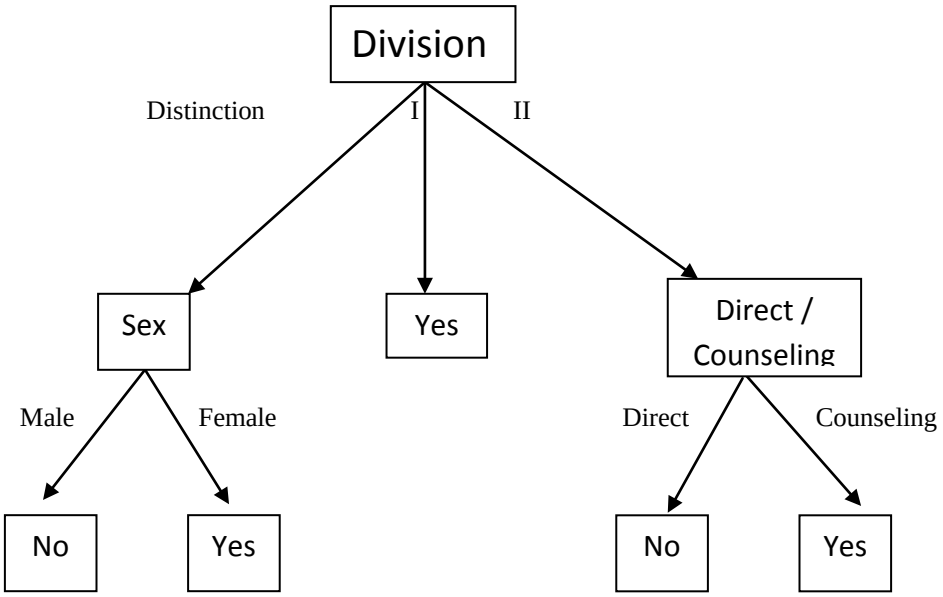


The above figure shows that tuples belonging to partition for division = I belong to same class. Since all tuples belong to class "yes", a leaf should therefore be created at end of this branch and labeled with "yes". The final decision tree created by algorithm is shown in Figure 2.

The rule set generated by ID3 clearly indicates that there is high probability of attrition in following cases (Figure 2)

1. When Division is I as all tuples belong to same class "yes".
2. If student has passed with distinction and is female.
3. If Division is II and has taken admission through counseling

FIGURE 2. DECISION TREE INDICATING WHETHER STUDENT IN INSTITUTION IS LIKELY TO WITHDRAW ADMISSION.



HANDLING PROBLEM OF ATTRITION OF STUDENTS

Institutions need to determine the extent of its own attrition problem, because of the complex nature of drop-in and drop-out patterns. Institute can be more helpful to students as they pursue their educational goals. To tackle attrition problem among students a committee comprising of senior faculty members and students need to be set up. This committee can have regular follow-up of freshly enrolled students and can apprise students about the facilities and latest developments regarding placements and cultural activities in institutes.

CONCLUSION

In this paper, a simple methodology based on decision tree is used to determine the set of students that are likely to leave institution after confirming the admission. This methodology will assist the academic planners to identify the set of students that are highly likely to attrite and devise the methods to minimize this rate of attrition among students. In fact analysis will assist the management to look into problems being faced by the students and take the corrective action to increase the retention rate among students.

REFERENCES

[1] Kova J. Zlatko, "Predicting student success by mining enrolment data ' Research in Higher Education Journal
[2] Arora Kumar Rakesh Kumar, Badal Dharmendra,"Evaluating students performance using k-means clustering",IJCST Volume IV , Issue 2
[3] Ramist L "College student Retention and Attrition – Research and Development", College Entrance Examination Board, New York
[4] Student Attrition:Consequences, Contributing factors and remedies,Ascend Learning,LLC,March 2012
[5] http://www.saedsayad.com/decision_tree.htm
[6] Han Jiawei, Kamber Micheline "Data Mining: Concepts and Techniques"Second Edition
[7] Arora Kumar Rakesh, Badal Dharmendra,"Location wise student admission analysis", International Journal of Computer Science, Information Technology and Security, Dec 2012.