EDUCATIONAL DATA MINING MODEL TO EVALUATE STUDENTS PERFORMANCE

Ch. Hrudaya Neeharika, Y.Md. Riyazuddin

hchitta@gitam.edu, rymd@gitam.edu

1Research Scholar, Dept of CSE, School of Technology, GITAM (Deemed to be University), Hyderabad
2Dept of CSE, Assistant Professor, School of Technology, GITAM (Deemed to be University), Hyderabad

Abstract: The present study aims to explain the ability of higher education technology to data mining by providing a university data mining model. The classification task in the analysis is used to evaluate the performance of the student, even though a decision-making process is used here since we used several methods for data classification. With this activity, we collect information identifying the performance of students during the final semester test. It helps recognize dropouts and students who need special care, so that they can provide the teacher with proper advice. In this study, we used Modified C4.5 to build a decision-tree as a proposed approach to a supervised learning model that involves two steps: the first step involves the discretionary consideration of every continuous attribute, rather than number values. The second step uses the average gain measure to choose the best feature instead of the gain ratio measure. The experimental findings reveal the total number of nodes without affecting the accuracy, with discretization technology showing an improvement of 4.58% as compared to the traditional C4.5 algorithm.

Keywords: Classification, Educational Data Mining (EDM), C4.5, Modified C4.5 algorithm.

I. INTRODUCTION

A country’s growth depends on its people's educational history. The high literacy rate is seen in most developing countries. There is also an increasing need for teaching and learning. As all government universities and colleges alone cannot satisfy this increasing demand, it has founded a significant number of private universities and colleges in the last decade, and the trend will increase in the future. In order to attain the desired development of any nation, the reasons behind lower higher education enrollments, weak teaching efficiency, and learning and study need to be uncovered. Many technologies and methods that can transform or improve the educational environment are now accessible [1-3]. The increasing digitalization of educational data has facilitated the collection of usable data and the extraction of relevant data to take correct decisions. Educational data provides considerable transparency, which has exploded our knowledge of education during the last decade. Education data is growing quickly as the educational system becomes more and more efficient. It has opened up new fields, such as immersive learning approaches and smart tools assisted by machines, and training games that have opened up the opportunity to gather and interpret student information and to understand patterns and trends in the results. They can aggregate data from online training programmes over a vast number of students and include several variables which can be explored for model construction through data mining algorithms [4-6].

In the data mining sector, recent phenomena in vast quantities of data are to be found. It has few uses for education, yet is widely used in business. Of course, it is possible to use data mining for the educational industry, for instance, to figure out which students will make significant contributions. From the pedagogical viewpoint, we are interested in mining models of students. Our aim is to identify how to use data mining strategies to mine, and
also how to explore and present models that are methodologically interesting for both
students and teachers. The primary goal of this study is to use some sort of mining
techniques to assess students’ success in different courses. In order to investigate student
results, data mining offers a large range of activities. In this analysis, the challenge of
evaluating the success of students is used, and the decision-tree system is used since there
are several methods used for classifying results. Information from the student management
framework was obtained to predict the results at the end of the semester, for example,
attendance, a class examination, a seminar, and a mark. This study assesses the performance
levels of students at school and university level with the help of modified decision tree
approaches.

II. BACKGROUND WORKS

In this subsection, some works on students’ performance with respect to datamining
techniques are discussed. In [7], the work is completely focused on socio-demographic and
students’ educational data that belongs to Indian universities. In addition, they built an
improved ID3-based decision-tab algorithm that will enable students to predict whether they
are going to continue or quit their studies. The Renyi entropy, the benefit of information, and
association functions improve the ID3 algorithm and the model developed from an efficient
decision tree that helps the superintendents of universities in adopting guiding principles for
increasing university enrollment and taking precautionary steps. The students who leave will
also find the causes and related factors. Their research findings show that the improved
algorithm of the decision-tree has greater consistency in education data than standard
classification algorithms. Their improved algorithm for choosing improved decision-making
factors and thus showed better classifiers with the proposed algorithm.

In [8], the authors presented the definition of the Augmented Intelligence (AUI) approach
to educational data mining (EDM). They adopted the AUI procedure in a mechanism that
produces a model predictive. An EDM end user can adjust it using a white-box learning
algorithm. The predictive adjustable model affects end-user experience and changes
predictive model results. The AUI approach produces new knowledge in the educational
environment when used in cycles. The probable educational data mining users built a
potential model in an AUI study and they monitored the evolution of models. The analysis
shows that the models generalize further over time and that the AUI approach avoids over-
fitting. In addition, the end-user understood that the cyclic structure of the AUI approach
allows for deeper data processing.

In [9], the authors determined the predictive importance of information, behavior, and
demographic attributes to the performance levels of the students studying at Alligarh School,
India. The factor analyzed the results of the various variables in order to determine the
predictive validity of these predictors by a smaller number of meaningful variables or
influences. Science Stream performance factors were established. In order to classify
variables that distinguish the predictors, they compared the prognostic benefits for top
performers and poor performers.

In [10], the authors adopted a system in which the academic achievement of students
reflects the history of academia. Prediction of performance supports educational institutes in
supporting decision-making and in developing better management strategies. For educational
institutions, this success record is significant since they will benefit from it by developing
their skills through knowledge of student outcomes. Data mining in education analyses these
data and extracts them from them. Then they decided on the status of the academic results.
They employed methods such as decision trees, sorting, data clustering, and neural networks,
and so on. In this study, they predicted the success of the students for the semester.

www.turkjphysiotherrehabil.org

4482
In [11], the authors constructed a technological-based cluster-pattern that is capable of illustrating its usage with two exemplary instances, which reflect the student's use of online learning by using a click-stream server. Cluster analysis may help researchers create profiles based on learning activities—including assignments and information sequence, time spent on a particular project or testing tools that are more essential for learning purposes. In a similar way, they use use-cases for analysis of the features of learning behavior, whereas learners take part in problem solving practice in an on-line learning context, using hierarchic or non-hierarchical based clustering approaches.

In [12], the authors provided an EDM application with many intelligence and other variables are illustrated (e.g., learning styles and personality types). The decision-tab model was applied in order to classify talented students, using student behavioral traits, different intelligences, and characteristics. They sampled 735 high school students. The built decision tree model with a validity of 70% shows that it may be workable to examine mathematically talented students utilizing techniques of data mining if basic characteristics are used.

III. PROPOSED MODEL

The activities shown in Figure 1 in education data mining utilizing cluster analytics and decision tree techniques are recognized according to the literature investigated in the past phase. In the current education framework, a student evaluation and final semester examination evaluate the success of a student. The teacher performs an internal assessment focused on the participation of students in educational events, including class tests, weekly assignments, skill programs, and lab work. In this current study, students in the final semester are examined and to complete a semester in the internal and final semester exams, each student must have minimum marks.
Data description:
This study collects data on the computer engineering department of course B. Tech from session 2016 to 2020 from the University of Hyderabad (Telangana). The data was initially 60 in size. As a preprocessing step, all the redundant data is removed and all the entries are listed in separate tables in one row.

Table 1: Student associated variables

<table>
<thead>
<tr>
<th>Variable type</th>
<th>Description</th>
<th>Expected Values</th>
</tr>
</thead>
</table>
| FSM           | First semester marks | (First > 70%  
                      Second >55 &<60%  
                      Third >46 &<55%  
                      Fail < 40%)     |
| SEM           | Semester performance| (Poor, Average, Good)                                |
| ASS           | Assignment         | (Yes, No)                                            |
| ATT           | Attendance         | (Poor, Average, Good)                                |
| PW            | Practical work     | (Yes, No)                                            |
| LSM           | Last semester marks| (First > 70%  
                      Second >55 &<60%  
                      Third >46 &<55%  |
Data selection and transformation: At this stage, only the fields needed for data mining were selected. Selected were certain derived variables. Although certain data is derived from the database for the variables. All the extracted variables in the database are listed in Table 1 for comparison purposes.

Decision Tree: The decision tree uses the representation tree to address the problem of matching class labels with each leaf node and reflects the attributes on the tree’s internal node. The decision tree begins with a root node where users will act. From this node, the decision tree algorithm divides recursively any node. Finally, the decision tree reflects a decision-making scenario and the results of any branch.

Classical C4.5 Decision tree Algorithm:

C4.5 follows ID3 and eliminates the constraint of categorical characteristics by defining dynamically (based on numerical variables) a discrete attribute which partitions the continuing attribute value into a discrete collection of intervals. C4.5 translates the trees that have been learned (i.e., ID3 algorithm output). We then assessed the accuracy of each rule in order to decide the order of operation. Take away a rule if the accuracy of the rule changes without it changing. If the rule is eliminated,

Entropy: It is the measurement of variance with a random variable and the impurity of a set of arbitrary instances is defined. The greater the information content, the greater the entropy.

\[
Entropy = \sum_j p_j \log_2 p_j
\]

Information Gain Ratio: When small or medium numbers of values are used, the information gain is good. For high values, it returns less information. The value of gain cannot achieve the total value of gain. C 4.5 employs a gain ratio by separating the activities depending on the test attributes to overcome these limitations. When selecting an attribute, the gain ratio considers the number and branch size. It corrects the information gain by taking into consideration the intrinsic information of a split.

Information gain = Entropy (S)- [Weighted Avg] * Entropy (each feature)

Gain (S, A) of an attribute A, relative to a collection of examples S.

Proposed Modified C4.5 model:

To improve the C4.5 based decision tree model, the concept of discretization approach has been incorporated into the attribute sequence splitting mechanism. Initially, this approach sorts data variables and segments them into equivalent bins. The following formula is used to discretize continuous values into lower values. A discrete symbol can dynamically distinguish which. The following is the equation used in the discrete method to optimize the C4.5 algorithm process.

\[
D(S) = \sum - (P_T) \log_2 (P_T)
\]
Where \( D \) is the entropy target, \( S \) is the split entropy base variable, \( \Sigma \) is the total number of variables and \( (P_i) \) is the maximum information. We based the sample calculations below with discretization methods on the proposed discretization method for the splitting of the continuous attribute.

**Mathematical analysis of the Goodness Score without involving the Discretization approach:**

We calculated the score of goodness by using the formula without including the discretization method:

\[
GS = \left( \frac{A + B}{C} \right) / T
\]

(4)

where \( GS \) corresponds to the goodness score, where \( A+B = \) the amount of time in branches and \( T \) to the number of branch levels.

**IV. RESULTS AND DISCUSSION**

We must first measure \( S \) entropy to calculate the information gain for \( A \) comparative to \( S \). This is a set of 60 instances of 25 ‘First’ and 15 ‘Second’, 11 "Third" and 9 "Fail". With no discretization, we performed the measurement of the Goodness Score using Eq. (4), with a result of 0.34.

![Figure 2: Online test category branches](image)

During three evaluation levels, the number of questions asked and answered are basic skills, technical and general knowledge, and then are separated into the overall instances in the dataset. The result is then again split into 3, which correlates to the online test category branches as seen in Figure 2.

**Calculation of Entropy**
Entropy(S) = \(-P_{first}\log_2(P_{first}) - P_{second}\log_2(P_{second}) - P_{third}\log_2(P_{third}) - P_{fail}\log_2(P_{fail})\)

\[
(5)
\]

\[
= -\left(\frac{25}{60}\right)\log_2\left(\frac{25}{60}\right) - \left(\frac{15}{60}\right)\log_2\left(\frac{15}{60}\right) - \left(\frac{11}{60}\right)\log_2\left(\frac{11}{60}\right) - \left(\frac{9}{60}\right)\log_2\left(\frac{9}{60}\right)
\]

= 1.673

Calculation of Information Gain

\[
Gain(S, FSM) = Entropy(S) - \left|\frac{S_{first}}{|S|}\right| - Entropy(S_{first})
- \left|\frac{S_{second}}{|S|}\right| - Entropy(S_{second}) - \left|\frac{S_{third}}{|S|}\right| - Entropy(S_{third})
- \left|\frac{S_{fail}}{|S|}\right| - Entropy(S_{fail})
\]

\[
(6)
\]

= 0.473

Mathematical analysis of Goodness Score computation without involving Discretization approach:

\[
GS = \frac{(20 + 15 + 18 + 12 + 30 + 05)}{3}/100
\]

GS = 0.33

Mathematical analysis of Goodness Score computation with involving Discretization approach:

**Basic skills**

D (answered) = 20/ (20+15) = 0.57

D (Unanswered) = 15/ (15+20) = 0.42

Score = E (20+15) = (0.57+0.42) + 0.57 \times \log_2 (0.57) - 0.42 \times \log_2 (0.42)

= (0.99) + 0.57 \times (-0.81) - 0.42 \times (-1.08)

= (0.99) + (0.46) + (0.45)

= 1.90

**Technical**

D (answered) = 18/ (18+12) = 0.60

D (Unanswered) = 12/ (12+18) = 0.40

Score = E (18+12) = (0.60+0.40) + 0.60 \times \log_2 (0.60) - 0.40 \times \log_2 (0.40)

= (1.00) + 0.60 \times (-1.73) - 0.40 \times (-1.32)

= (1.00) + (1.03) + (0.52)
After analyzing the combined population of graduate level (GL) students, Figure 3 and Table 2 show the drop-out rates for the three different types of skills. According to this graph, there are two types of drops: term drop percent and cumulative drop percent. As can be seen in Figure 3, the majority of student dropouts occur within the first four years of school. There were two dropouts in the last year, where a cutoff of 50% is used in this study. Students’ technical abilities are the most often lacking, according to the data.

Figure 3: Drop-out rates for the three different types of skills

<table>
<thead>
<tr>
<th>Table 2: Dropout of students in four years</th>
</tr>
</thead>
<tbody>
<tr>
<td>First year</td>
</tr>
<tr>
<td>Basic skills</td>
</tr>
<tr>
<td>Technical skills</td>
</tr>
<tr>
<td>General knowledge</td>
</tr>
</tbody>
</table>

V. CONCLUSION
In this paper, we discuss the main elements that are needed to construct an educational datamining tool. Because of large data stored in databases with information on the actions of students in e-courses of higher education institutions, this field has become popular in recent years. The integration of discretization approach provided 1.51 value compared with the traditional C4.5 model, whereas a value of 0.33 resulted by means of the modified C4.5 model in online skills testing datasets comprising three stages of evaluation: basic, technical, and general knowledge.

REFERENCES


