Web usage Mining Application for Enrollment Management in Educational Institutes

Anand Gandhe¹, Dr. B.Dhanasekaran²
Associate Professor¹, Principal²
School of Computers, IPS Academy, Indore, India¹
Tamilnadu, India²

Abstract:
The rates of enrollments for higher education has increased but not as much as the number of higher institutions are increasing. Hence it has become important to understand the requirement of students and their academic progression. Educational Data Mining helps in a big way to answer the issues of predictions and profiling of not only students but other stake holders of education sectors. Web data is to be processed to achieve customer needs and satisfaction, because the web data is unstructured or semi-structured, one cannot apply data mining techniques directly on such data, this is where Web Mining has been evolved. Web mining is an interesting discipline in the domain of data mining where information mining strategies are utilized for extracting data from the web servers. Web Mining is the way towards applying information mining strategies on web data to automatically and quickly extract useful information and also to discover interesting patterns. Web usage mining uses the methods which can predict the user behavior. Web usage mining is the procedure of extracting data or knowledge that is valuable use designs from web information, to understand the user's needs and implement them in the web applications. Web usage mining is used to analyze the web logs and site files, personal information of learners, learning results, learning behavior of the users. This study explains identifying at-risk students, selection of course by the student, identifying priority learning needs for different groups of students.

Keywords: web mining; web usage mining; pre-processing; pattern discovery; pattern analysis; applications.

1. INTRODUCTION

The development of any country depends upon the educational background of its citizens. Growing digitization of educational data has helped researcher to easily capture these available data and extract meaningful information to take corrective decisions. Educators are responsible for designing, planning, building and maintaining educational systems, while students use and interact with them. The application of data mining is different for educators and students. For students, the goal is to discover activities, resources and learning tasks that improve their learning, based on their attitude and likings while for educators, the goal is to have more feedback from students for evaluating the structure of the course content and its effectiveness on the learning process to classify students based on their needs to discover information to improve the adaptation and customization of the course, etc. Web Mining is the way toward applying information mining strategies on web data to automatically and quickly extract useful information and also to discover interesting patterns. Although web mining has its establishes profoundly in information mining, it isn't same as information mining. The unpredictability of web mining depends on the unstructured nature of web data. Web mining entail the analysis of a particular web sites server logs basically called web server logs, which contain the complete interaction list of a particular user when accessing the web site. The knowledgeable information from such analysis from server logs is very helpful in almost all the web applications. Web mining is separated into three categories, web content mining: is the procedure of extricating data or knowledge from the content of web documents such as text and images, web structure mining: is the procedure of extricating data or knowledge from the structure of web pages and impact of this structure on traversal through these web pages, web usage mining is the procedure of extricating data or knowledge i.e. valuable use designs from web information, to understand the user's needs and implement them in the web applications. The vital uses of web use mining are personalization, web site design, ecommerce, web advertising or web marketing, fraud detection, transaction analysis, educational data mining etc. The main objective of this paper is to use data mining methodologies to select student for enrollment in a particular course. Data mining provides many tasks that could be used to select good student for a particular course. In this research, the classification task is used to evaluate previous student’s performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information like stream, marks in graduation, students performance etc. were collected from the student’s management system, to predict the suitable student for enrollment in a particular course.

2. 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. Decision tree are commonly used for gaining information for the purpose of decision –making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each
node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. The four widely used decision tree learning algorithms are: ID3, CART, CHAID and C4.5.

1. **ID3** algorithm introduced by J. R. Quinlan is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes. The attribute with the highest information gain or greatest entropy reduction is chosen as the test attribute for the current node.

2. **C4.5** is an algorithm used to generate a decision tree developed by Ross Quinlan. C.4.5 is a successor of ID3. C.4.5 made a number of improvements to ID3. C.4.5 uses Gain ratio as an attribute selection measure. Also C.4.5 can handle both discrete and continuous attribute.

3. **CART** (Classification And Regression Tree) algorithm, which was proposed by Breiman, is conceptually the same as that of ID3. The impurity measure used in selecting the variable in CART is Gini index. If the target variable is nominal it generates classification tree and for continuous-valued numerical target variable it generates regression tree.

4. **CHAID** (CHi-squared Automatic Interaction Detector) uses Chi square contingency test for tree construction in two ways. First, it determines whether levels in the predictor can be merged together. Once all predictor level is compressed to their smallest significant form, it determines most significant predictor in distinguishing among the dependent variable levels.

3. **THE ID3 DECISION TREE**

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan. 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 ID3 Algorithm

ID3 (Examples, Target Attribute, Attributes)

- Create a root node for the tree
- If all examples are positive, Return the single-node tree Root, with label = +.
- If all examples are negative, Return the single-node tree Root, with label = -.
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
- Otherwise Begin
  - A = The Attribute that best classifies examples.
  - Decision Tree attribute for Root = A.
  - For each possible value, vi, of A,
    - Add a new tree branch below Root, corresponding to the test A = vi.
    - Let Examples(vi) be the subset of examples that have the value vi for A.
    - If Examples(vi) is empty
      - Then below this new branch add a leaf node with label = most common target value in the examples.
    - Else below this new branch add the sub tree ID3 (Examples(vi),Target_ Attribute, Attributes – {A})
  - End
- Return Root

4. **DATA SELECTION AND PREPROCESSING**

Master of Computer Application (MCA) is a Three Year (Six Semester) full time course. The MCA course is open to graduate in any discipline with Mathematics as a subject at 10+2 level. That candidate, who has done BCA and B.Sc. Computer Science are exempted from mathematics at 10 + 2 level. Another Course is MCA Lateral Entry is a two years course after graduations. Candidates who have completed either BCA or B.Sc. Computer Science are given admission to this course. Further, there are two courses BCA and BCA+MCA (Dual Degree) in which criteria is students with mathematics at 10+2 level can take admission in these courses. The data was collected through the enrolment form filled by the student at the time of admission on the website. Students enter their demographic data (category, gender etc), past performance data (SSC or 10th marks, HSC or 10 + 2 exam marks and Graduation Marks etc.), address and contact number. From these the attributes that possibly influence their result are selected as shown in Table 1. Most of the attributes reveal the performance of the students.

![Table 1. Student Related Variables](http://ijesc.org/)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Students Sex</td>
<td>[Male, Female]</td>
</tr>
<tr>
<td>Cat</td>
<td>Students category</td>
<td>[General, OBC, SC, ST]</td>
</tr>
<tr>
<td>GSS</td>
<td>Students grade in Senior Secondary education</td>
<td>{O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%}</td>
</tr>
<tr>
<td>GMSS</td>
<td>Students grade in Math at Senior Secondary education</td>
<td>{O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%}</td>
</tr>
</tbody>
</table>
5. MODEL CONSTRUCTION

Rapid Miner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures. The general features of rapid miner are an environment for machine learning and data mining processes. It represents a new approach to design even very complicated problems by using a modular operator concept which allows design of complex nested operator chains for huge number of learning problems. Rapid miner uses XML to describe the operator trees modeling knowledge discovery process. It has flexible operators for data input and output file formats. It contains more than 100 learning schemes for regression classification and clustering analysis. Rapid miner supports about twenty two file formats. Rapid Miner has a lot of functionality, is polished and has good connectivity. Rapid Miner includes many learning algorithms from WEKA. It is solid and complete package. It easily reads and writes Excel files and different databases. We can program by piping components together in a graphic ETL work flows. If you set up an illegal work flows Rapid Miner suggest Quick Fixes to make it legal.

Advantages:

- It has the full facility for model evaluation using cross validation and independent validation sets.
- Over 1,500 methods for data integration, data transformation, analysis and, modeling as well as visualization – no other solution on the market offers more procedures and therefore more possibilities of defining the optimal analysis processes.
- Rapid Miner offers numerous procedures, especially in the area of attribute selection and for outlier detection, which no other solution offers.

6. RESULTS OBTAINED

The ID3 decision tree generated. The accuracy of the model is 66.66%. That is out of 500 instances 330 instances are correctly classified. The most important attribute in predicting student’s enrollment is found to be GS. The social attributes like category, Medium, College Location and Admission Type are not appearing in the decision tree indicating less relevance of the prediction with such attributes. The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules in Table 2.

---

Table 2. Rule Set generated by ID3

<table>
<thead>
<tr>
<th>Rules for Enrolment in BCA+MCA Dual Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF GSS ≥ ‘C’ and GMSS ≥ ‘C’ THEN Enrolment = ‘BCA+MCA Dual Degree’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rules for Enrolment in BCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF GSS = ‘D’ and GMSS = ‘D’ and GSS = ‘E’ and GMSS = ‘E’ THEN Enrolment = ‘BCA’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rules for Enrolment in MCA Lateral Entry Two Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF GS = ‘BCA or B.Sc. Computer Science’ and GMSS ≥ ‘C’ THEN Enrolment = ‘MCA Lateral Entry Two Years’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rules for Enrolment in MCA Three Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF GS = ‘Any Graduate with Math’ and GMSS ≥ ‘C’ Enrolment = ‘MCA Three Years’</td>
</tr>
</tbody>
</table>

The classification matrix has been presented in Table 3, which compared the actual and predicted classifications.

---

Table 3. Classification Matrix-ID3 prediction Model

<table>
<thead>
<tr>
<th>Actual Students Enrolment</th>
<th>Predicted</th>
<th>% of Correct Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCA Lateral Entry Two Years</td>
<td>120</td>
<td>66.66%</td>
</tr>
<tr>
<td>MCA Three Years</td>
<td>36</td>
<td>60%</td>
</tr>
<tr>
<td>BCA+MCA Dual Degree</td>
<td>45</td>
<td>75%</td>
</tr>
<tr>
<td>BCA</td>
<td>80</td>
<td>66.66%</td>
</tr>
</tbody>
</table>

---
7. CONCLUSION

The important attribute in predicting students’ enrolment is found to be Graduation Stream (GS). The study shows that the student past academic performance can be used to create a decision tree model using ID3 algorithm that is used for predicting Student’s enrollment in MCA course and other related courses. The study found that students with background of mathematics or computers are more suitable in MCA course than students with other backgrounds.

8. REFERENCES


[5]. B Bhavani, Dr. V. Sucharita & Dr. K.V.V. Satyanarana, Review on Techniques and Applications Involved in Web Usage Mining, International Journal of Applied Engineering Research ISSN (0973-4562 Volume 12, Number 24 (2017) pp. 15994-15998


