A Maximally Specific Hypothesis for Predicting the Employability Requirements in Higher Educational Institutions

G. Sriram
School of Distance Education
Andhra University
Visakhapatnam, India
E-mail: ram.gopalam@gmail.com

Y. Srinivas
Department of Information Technology
GIT, GITAM University
Visakhapatnam, India
E-mail: drysr@gitam.edu

K. Thammi Reddy
Department of Computer Science & Engineering
GIT, GITAM University
Visakhapatnam, India
E-mail: thammireddy@gitam.edu

Abstract: Educational Data mining is becoming one of the fascinating areas in the field of Data Mining by various research communities. In India the academic institutions are now adopting ICT initiatives for the academic functionalities like teaching learning, evaluation, admissions etc. aimed at collating, storing and viewing of the data. This has paved way in the creation of student data repositories, which can be used for educational data analysis. This paper aims at reducing the representation of rules using Maximally Specific Hypothesis for predicting the attributes which play a vital role in the employability of students. The results are compared with standard decision tree algorithms like ID3 and Random Forest and achieved the best results in representation as well as reduction of employability attributes.

Keywords: Educational Data Mining, Inductive Learning, Concept Learning, Data Mining.

INTRODUCTION

Educational Data mining is becoming one of the fascinating areas in the field of Data Mining by various research communities [16]. Data Mining is a process of extracting knowledge from huge repositories. It is used to uncover patterns which are interesting and can be very much useful for academic environments to derive the objectives of the knowledge acquiring process. Data mining provides rich set of algorithms for various tasks using which one can extract golden nuggets. [6], [7].

In India the academic institutions are now adopting ICT initiatives for the academic functionalities in teaching learning, evaluation, admissions etc. aimed at collating, storing and viewing of the data. Many applications to perform these activities are now available either as an open source or proprietary based. This situation has paved the way for crunching huge amount of data which is in the institutional repositories for uncovering knowledge to improve their performance.

In the process of acquiring knowledge the steps involved are 1.Content dissemination 2. Content Understanding, 3. Evaluation of the learner 4. Content Revision, 5. Employability, 6. Feedback Mechanism. The measure of knowledge from the above process is mainly aimed on learners, whose capabilities are either measured directly or indirectly using various factors like employability, scoring marks, etc. Recently many Higher Educational Intuitions are measuring their student employability as one of the yard stick in knowledge acquiring process. True that the employers are also focused on evaluating skills acquired in their academic progression [1], [2].

I. RELATED WORK

Educational Data Mining has attracted many researchers in recent times. One of the open academic initiatives by Sakai is a community of academic institutions, commercial organizations and individuals who work together to developed a common Collaboration and Learning Environment [11]. It is an educational software platform distributed under the Educational Community License. Sakai is used for teaching, research and collaboration. Systems of this type are also known as Course Management Systems (CMS), Learning Management Systems (LMS), or Virtual Learning Environments (VLE). P. Ajith, et al [10] focused on Academic Data Mining and used classification techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor. The discovered knowledge can be used for prediction.
and analysis purposes of student patterns. M. Ramaswami [9] presented a methodology to investigate the most relevant subset features with minimum cardinality for achieving high predictive performance by adopting various filtered feature selection techniques in data mining but also to evaluate the goodness of subsets with different cardinalities and the quality of six filtered feature selection algorithms in terms of F-measure value and Receiver Operating Characteristics (ROC) value, generated by the NaïveBayes algorithm as base-line classifier method. Wilairat Yathonchhai [16] have considered factors affecting student’s dropout rate. These factors are conditions related to the students before admission, factors related to the students during the study periods in the university, and all factors including the target value to be predicted for factors analysis. They used tree-based classification algorithm, J48 or C4.5, and Naïve Bayes algorithm to analyze the data [5], [6], [8].

The earlier work is focused on analyzing independent components of the knowledge acquiring process [3], [4]. However, in order to evaluate the learning abilities and employability skills of the learner, there is a great need to develop a system which integrates all the aspects of knowledge acquisition [12]. This helps in identifying the gap and formulate a link analysis to further strengthen the learning and employable abilities of the learner.

1. To identify the Attributes of importance.
2. To Generate Inductive learning hypothesis.
3. Inductive learning hypothesis vs Decision Tree Classification.

II. METHODOLOGY

The proposed paper aims at developing a system using inductive learning process for generating the hypothesis or crisp rules on a real time student database of 771 instances. These rules can be used to identify the key attributes that is required for employability among the graduates, which can be used as predictor attribute for employability. The earlier methods were depended on classification the generated rules are either too generic or too specific; if the rules are more it is time consuming else if rules are few we may not classify the data. Generating an optimum number of rules is becoming a challenging task. In particular application relating to educational mining requires crisp generation of rules. The real time student database consists of 17 attributes to identify the attributes of importance an attribute oriented Induction (AOI) is used to reduce it to six attributes. The method is compared with the decision tree classification.

III. DATA PRE PROCESSING

The data pre processing is an important activity in identifying the attributes of relevance and also removing noise contained in the real time data set.

An Attribute Oriented Induction (AOI) method is used to reduce the representation of the attributes [8]. It is based on data generalization and summarization which involves various steps in identifying attributes of importance. The AOI is applied on the attribute set to remove an attribute which of less importance or an attribute may be generalized.

An attribute will be removed if an attribute contains distinct values or its next level concepts are expressed in terms of other attributes. The working of the algorithm is given below.

i.) The task-relevant data is represented as the initial working relation.

ii.) Attribute-removal: remove attribute A if there is a large set of distinct values for A but (1) there is no generalization operator on A, or (2) A’s higher level concepts are expressed in terms of other attributes

iii.) Attribute-generalization: If there is a large set of distinct values for A, and there exists a set of generalization operators on A, then select an operator and generalize A

iv.) Attribute-threshold control and generalized relation threshold control.

The attributes are: ID Number, Name, DOB, SSC, Inter, B.Tech, Mobile, email, Pass category, Social category, Parents Name, Parent’s Mobile, Permanent Home Town/Nearest Town, Permanent Address, Alternate Contact Number, Alternate email ID, Gender. After applying the AOI algorithm the reduced set of Attributes: Age, SSC, Inter, B.Tech, Gender, Status, and Count.

IV. PROPOSED CONCEPT LEARNING APPROACH

Inferring automatically a concept from the available data is concept learning. The concept learning is also used for searching through a large space of hypothesis [13]. The work is aimed at generating the best representational hypothesis...
from the available student data further which can be used as search mechanism to derive the knowledge.

Since the database consists of huge search space there is a need to generate a lesser dimension space using concept learning approach using the Maximally Specific Hypothesis. From the student placement data most specific possible hypothesis are generated, then generalize this hypothesis each time it fails to cover an observed positive training example.

1. Initialize h to the most specific hypothesis in H
2. For each positive training instance x
   a.) For each attribute constraint ai in h
      If the constraint ai is satisfied by x
         Then do nothing
      Else replace ai in h by the next more general constraint that is satisfied by x.
3. Output hypothesis h.

V. EXPERIMENT ANALYSIS

The algorithm was applied on the student database of 771 instances with 3 attributes by partitioning based on the branch on which one can analyze the .

The output generated by applying the above algorithm on the Civil branch sample data is present.

The Hypothesis are:
1. < ?, ?, D > < 16 selected, 39 Not Selected>
2. < ?, ?, F > < 6 selected, 0 Not Selected>
Similarly the Hypothesis for the other branches are also derived.

CSE branch:
1. < ?, ?, D > < 124 selected, 52 Not Selected>
2. < ?, ?, F > < 27 selected, 57 Not Selected>
ECE branch:
1. < ?, ?, D > < 189 selected, 01 Not Selected>
2. < ?, ?, F > < 18 selected, 60 Not Selected>
3. < ?, ?, S > < 6 selected, 54 Not Selected>

The hypothesis that is generated is compared with standard algorithms like ID3 and Random Forest. [14], [15].

<table>
<thead>
<tr>
<th>Branch</th>
<th>Size of the dataset</th>
<th>No. of rules generated from ID3</th>
<th>Reduced Set using Maximally Specific Hypothesis</th>
<th>% reduced rep. from original data</th>
<th>% reduced in rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civil</td>
<td>60</td>
<td>19</td>
<td>2</td>
<td>32</td>
<td>11</td>
</tr>
<tr>
<td>CSE</td>
<td>254</td>
<td>40</td>
<td>2</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>ECE</td>
<td>328</td>
<td>28</td>
<td>3</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>MEC</td>
<td>129</td>
<td>31</td>
<td>3</td>
<td>24</td>
<td>10</td>
</tr>
</tbody>
</table>

From the above hypothesis it is observed that for the branches of CSE , ECE and Mech. Engg. the present pursuing degree plays a vital role in the placement, where as for civil dept. this consideration may not be true.

VI. CONCLUSION:

The results of this proposed method clearly shown that the Maximally Specific Hypothesis is well suited for educational data mining. These hypotheses can be used to identify the key attributes that is required for employability among the graduates, which can be used as predictor attribute. Using AOI method we can able to reduce the representation of attributes set there by identifying the attributes of important. It also paved a way for data reduction technique. The results clearly show that we can able to reduce the rule representation to 11%.

REFERENCES

conference on Computing and Control Engineering (ICCCE 2012), 12&13 April, 2012


