STUDENTS’ PERFORMANCE PREDICTION USING GENETIC ALGORITHM

Ruhi R. Kabra¹ and R. S. Bichkar²

¹Department of Computer Engineering, G. H. R. College of Engineering and Management Ahmednagar, India
²Department of Computer Engineering, G. H. R. College of Engineering and Management, Pune, India

Corresponding author: Email:ruhi.kabra@raisoni.net

ABSTRACT:

Decision tree models are commonly used in educational data mining to examine the data and induce a tree that will be used to make predictions about educational data. This study enables to obtain the decision tree models that predict the academic performance of the engineering students in contact education system. Genetic algorithm is a powerful search and optimization technique that has shown promise in obtaining good decision trees. Decision trees are evolved using greedy as well as evolutionary algorithms. The results are discussed with respect to the accuracy and size of the tree induced using genetic algorithm and J48 (from WEKA). Also the attributes that are important for prediction of First Year engineering students results are also identified.

Keywords: Educational Data Mining, Decision Trees, Genetic Algorithm.

[1] INTRODUCTION

The technique used for prediction of engineering students’ result is classification using decision trees. Decision tree induction algorithms present several advantages over other learning algorithms, such as robustness to noise, low computational cost for generating the model, and ability to deal with redundant attributes. Besides, decision trees are simple to interpret. On the other hand, most decision tree induction algorithms are based on a greedy top-down recursive partitioning strategy for tree growth. One major drawback of greedy search is that it usually leads to sub-optimal solutions. Hence, other approach that has been used is the induction of decision trees through Genetic Algorithms. Instead of local search, GAs perform a robust global search in the space of candidate solutions. As a result, GAs tend to cope better with attribute interactions than greedy methods.

[2] BACKGROUND

2.1. Decision Trees
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 [4]. 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.

A decision tree is constructed from a training set, which consists of data tuples. Each tuple is completely described by a set of attributes and a class label. Attributes can have discrete or continuous values. Decision trees are used to classify the data tuples whose class label is unknown. Based on the attribute values of the tuple, the path from root to a leaf can be followed. The class of the leaf is the class predicted by decision tree for that tuple.

The task of constructing a tree from the training set has been called tree induction or tree building. Most existing tree induction systems adopt a greedy (i.e. non-backtracking) top-down divide and conquer manner where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. Recent developments suggest the use of genetic algorithms to avoid local optimal decisions and search the decision tree space with little a priori bias.

2.2. Genetic Algorithm

Genetic Algorithms are search algorithms that are based on concepts of natural selection and natural genetics as explained in [1], [2]. Genetic algorithm incorporates the processes observed in natural evolution. The genetic algorithm is different from other search methods in a way that it searches among a population of points. It works with a coding of parameter set instead of the parameter values themselves. It also uses objective function information without any gradient information. GAs efficiently search the irregular space and therefore they are applied to a variety of function optimization, parameter estimation and machine learning applications. The framework of genetic algorithm is as follows

1. Formulate initial population
2. Randomly initialize population
3. Repeat
4. Evaluate objective function
5. Find fitness function
6. Apply genetic operators
   (a) reproduction
   (b) crossover
   (c) mutation
7. Until stopping criteria

Decision trees can be evolved using genetic algorithm because we can use a tree structure to represent decision trees and the mutation-crossover operators can be efficiently altered to match this structure.

[3] LITERATURE REVIEW
Romero et al. [7] tested genetic algorithms on the Web-based Hypermedia Course and they show that genetic algorithm is a good alternative for extracting a small set of comprehensible rules.

Kalles and Pierrakeas [8] have analyzed students’ academic performance throughout the academic year, as measured by the homework assignments, attempted to derive short rules that explain and predict success or failure in the final exams using genetic algorithm based induction of decision trees.

Kalles and Xenos [9] used combination of genetic algorithm and decision trees (GATREE) on students data (at HOU) to suggest a quality control system in an educational context.

J. Bala et al.[10] used GA to search the space of all possible subsets of large set of features. For a given subset a decision tree is generated using ID3. The classification performance of the decision tree on unseen data is used as the measure of fitness for the given feature set, which in turn is used by GA to evolve better feature sets. The process is repeated until a feature subset is found with satisfactory classification performance.

Bhardwaj and Pal [11] used Bayes classification technique to performance of BCA students (UP, India). Most of the feature selected focus on socio-economic background of the student. It was found that the factors like students’ grade in SSC, living location, medium of teaching, mother’s qualification, students’ other habits, family annual income and students family status were highly correlated with student academic performance.

Akinola, Akinkunmi, Alo [12] used ANN backpropagation algorithm is used on the sample data of computer science students( University of Ibadam , Nigeria). Results show that candidates with good background in physics and mathematics will perform efficiently in computer programming and the pre-higher institution qualification would contribute immensely to the performance of students in their chosen course of studies.

Bresfelean worked on the data collected through the surveys from senior undergraduate students at the faculty of economics Business administration in Cluj-Napoca [13].Decision tree algorithms in the WEKA tool, ID3 and J48 were applied to predict which students are likely to continue their education with the postgraduate degree. The model was applied on two different specializations students’ data and an accuracy of 88.68.

S. Ghosh et. al. [14] used genetic algorithm to find all the frequent itemsets from given data sets.

R. Barros et.al,[15] presented the survey of evolutionary algorithms like genetic algorithm and genetic programming and reviewed applications of evolutionary algorithms for decision tree induction in different domains, such as software estimation, software modules protection and cardiac imaging data. Advantages and drawbacks of decision tree induction using evolutionary algorithms are also discussed along with the discussion of objective function, crossover and mutation operator selection, parameters setting for the same.

[4] GENETIC ALGORITHM FOR DATA MINING

The flowchart to illustrate the use of GA to predict performance of students is shown in [Figure 1].

4.1. Feature Selection
First it is important to identify the features that are going to affect students’ result. The literature survey shows that, researchers have considered combination of different attributes like students’ social background, economic conditions, family details, performance in the past exam. It is also observed that the features that affect may vary for different countries, different social and educational environment also. So the attributes that possibly influence their result are selected. The selected attributes are branch of engineering, SSC marks (math, science and aggregate percentage), SSC board, HSC marks (Math, PCM – Physics, Chemistry, Math marks, aggregate percentage and Common Entrance Test marks), Gender, living location, category. Most of the attributes reveal the past performance of the students. Reason behind concentrating on the past performance data is
1. Data is available in the administrative department of the institute.
2. If student has performed well in the past, it is most likely that he will perform well in subsequent exams as well.
3. It is important to concentrate on only the data that is available with correct values and highly influence the result.

The data of First year engineering students of Pune University was collected in Excel sheet and then stored in student.arff file. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

Figure: 1. Data mining using GA
4.2. Data Collection and Data Preprocessing

Data of 346 students of the engineering institute are collected who appeared for the first year of engineering in the year 2009-10, 2010-11. The data was collected through the enrolment form filled by the student at the time of admission. The student enters their demographic data (category, gender etc.), past performance data (SSC or 10th marks, HSC or 10+2 exam marks etc.), address and contact number.

The collected student data are preprocessed. The data are categorized according to values that an attribute can take. For example, the SSCpercent can have one of the values as Distinction (above 75), First class (60 to 75), Higher Second class (50 – 60), Second class (below 50). The categorized data is stored in ARFF format.

4.3. Apply GA

The Genetic Algorithm is applied to induce decision tree. Initial population of n-ary trees is created. GA performs selection, crossover and mutation operation followed by evaluation as long as stopping criteria is not satisfied (usually number of prespecified generations). The number of generations and population size can be set through parameters. The trees are created randomly. Any of the attribute in the ARFF file can be selected and added to the tree as a root or internal node. The leaf node is FE Result and can take the value as PASS, FAIL or ATKT (in three class prediction) and Promoted (PASS and ATKT) or FAIL (in two class prediction).

The fitness function gives the goodness of the tree. For each tree, its fitness function is calculated. Accuracy of the tree is selected as fitness function. To calculate the accuracy, the data part of the ARFF file is scanned.

The population of individuals created with the crossover and mutation operators is merged with the previous population and the worst individuals are removed to return to the original population size.

The crossover method is used by the genetic algorithm to mate individuals from the population to form new offspring. Sexual crossover takes four arguments: two parents and two children. If one child is nil, the operator is able to generate a single child. The crossover operator chooses two random nodes and just swaps those nodes’ sub-trees as shown in [Figure 2].

The mutation operator is defined as destructive mutator which destroys a subtree from the chosen individual tree. Mutation operator may result invalid decision tree where the leaf node is not bearing class label. In such case a node with random class label is attached at
mutating point. Mutation is not applied at leaf node as the leaf node is bearing class label. [Figure 3] shows mutation operator.

Each path traversed from root to leaf yields a rule. Such rules are extracted to create the model. Because of the genetic operators it is possible that two identical path appear from leaf to root. Such individuals are identified and one of such duplicate paths is removed from the tree. Because of this no duplicate rules are generated in the model.

After the termination of GA, the best decision tree obtained is used for prediction of students’ exam result.

Figure: 2. Crossover

Figure: 3. Mutation

[5] RESULTS
Decision trees are induced using genetic algorithm using GAlib. GAlib is a C++ library developed by Matthew Wall [16] designed to assist in the development of genetic algorithm applications. The library contains numerous classes that other functionality and ability in the design of optimization applications with genetic algorithms. This library was programmed so that it may be used on a variety of compilers on many platforms. A new genome class is created by multiply inheriting from the base genome class. The initialization method and operators are defined which are used by the genetic algorithm defined in the library.

The First Year student dataset is used for training. All the attributes are descretized. First the initial populations of n-ary decision trees are created and then crossover and mutation is applied for number of generations. The best individual is found and considered as resultant model. The crossover probability is 0.6 and the mutation probability is 0.01. The best individual is represented in the form of if then rules. The prediction model for three class prediction (i.e. PASS/FAIL/ATKT) and prediction model for two class prediction (i.e. PASS/FAIL) are obtained. Similarly an ARFF file of students’ Mathematics I result data is created. All other attributes are same except the target variable. The model for prediction of Mathematics I result is created. These models show that HSC an SSC marks are very important in prediction of FE result. Other attributes like category, living location, gender have less appearance in the models and do not play important role in prediction of FE result.

The example decision tree induced using J48 from WEKA on the same dataset [17] shown in [Figure 4].The accuracy of this model is 69%, that is out of 346 instances 242 are correctly classified. The important attributes identified are HSC CET marks, board at secondary level, science marks in SSC exams, PCM marks in HSC.
Figure: 4. Decision trees induced using J48

[Figure 5] shows the decision tree induced using GA for two class prediction of FE result. The important attributes for the prediction are HSC percentage and HSC CET. The students who got distinction in HSC are promoted. The students getting First Class in HSC and getting good CET marks (B or C grade, A grade samples are not many in training data) are likely to pass, but the same with low CET marks (D grade i.e. less than 80) are likely to fail. The students with HSC percent Second class are likely to fail. The accuracy is 64 % with tree size as 7 nodes.

Figure: 5. Decision trees induced using GA
These trees are compared with the trees induced using the genetic algorithm with respect to their size and accuracy as shown in Table 1. Ideally the tree should be accurate as well as small in size. The table shows the comparison. The above discussion shows that the GA induced trees observe the accuracy slightly less than J48. However GA is a powerful optimization technique and it is quite possible to obtain further improvements in result by using different GA parameters and GA types. Authors are currently exploring these possibilities.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>J48</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Undertaken</td>
<td>Accuracy</td>
<td>Size</td>
</tr>
<tr>
<td>FE Result Three class prediction</td>
<td>60</td>
<td>11</td>
</tr>
<tr>
<td>FE Result Two class prediction</td>
<td>69</td>
<td>9</td>
</tr>
<tr>
<td>Mathematics I result prediction</td>
<td>69</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 1. Comparison of GA and J48 induced trees

[6] CONCLUSION

Decision trees can be effectively used for predicting the result of engineering students. Decision trees can be induced using greedy algorithms as well as evolutionary algorithms. It is observed that the accuracies of early prediction are in the range from 59% to 69%. The attributes describing student’s past performance in various examinations play important role in first year engineering students’ result prediction. Although the accuracies are not very high, the obtained values are quite acceptable as we get good indication about result of forthcoming exam well in time and can be used to give additional inputs to students. The results are sensible to the type of students and academic input.

REFERENCES

Author[s] brief Introduction

Ruhi Kabra

Ruhi Kabra obtained her BE in Computer Science and Engineering from SGGS Institute of Engineering and Technology, Nanded, and ME from G. H. Raisoni College of Engineering and Management, Pune. Her research interests include Business intelligence and data mining

R. S. Bichkar
R S Bichkar obtained his BE and ME degrees in electronics from the SGGS Institute of Engineering and Technology, Nanded, 1986 and 1990 respectively, and his PhD from IIT Kharagpur in 2000. He is presently a professor in the Department of Electronics and Telecommunication Engineering, G H Raisoni College of Engineering and Management, Pune. His research interests include application of genetic algorithms to various search and optimization problems in electronics and computer science.