DATA MINING METHODOLOGIES TO STUDY STUDENT'S ACADEMIC PERFORMANCE USING THE C4.5 ALGORITHM

Hythem Hashim¹, Ahmed A. Talab², Ali Satty³, and Samani A. Talab¹

¹ Faculty of Computer Science and Technology, Alneelain University, Khartoum, Sudan.
² White Nile College for Science and Technology, White Nile State, Kosti.
³ School of Statistics and Actuarial Sciences, Alneelain University, Khartoum, Sudan.

ABSTRACT

The study placed a particular emphasis on the so called data mining algorithms, but focuses the bulk of attention on the C4.5 algorithm. Each educational institution, in general, aims to present a high quality of education. This depends upon predicting the students with poor results prior they entering in to final examination. Data mining techniques give many tasks that could be used to investigate the students' performance. The main objective of this paper is to build a classification model that can be used to improve the students’ academic records in Faculty of Mathematical Science and Statistics. This model has been done using the C4.5 algorithm as it is a well-known, commonly used data mining technique. The importance of this study is that predicting student performance is useful in many different settings. Data from the previous students' academic records in the faculty have been used to illustrate the considered algorithm in order to build our classification model.

KEYWORDS

Data mining, The C4.5 algorithm, Prediction, Classification algorithms.

1. INTRODUCTION

The main objective of Faculty of Mathematical Science and Statistics in Alneelain University is to give a good quality education to the students as well as to develop the issues concerning quality of decisions with respect to managerial matters. Thus, one recommendation is detect knowledge from educational records to study the main attributes that may affect the student's performance in the considered faculty. This can be considered as an important and helpful aspect in choosing the right decisions to improve the quality of education. As well, it helps the academic staff in the faculty in order to support their decision making process with respect to the following aspects: (1) improve the student's performance; (2) improve teaching; (3) minimize the failure rate; and (4) other benefits. Data mining analysis is a good option to achieve the aforementioned objective as it gives many tasks that could be used to investigate the student performance.
When there is a need to data mining, the following question is forced upon researchers: What is data mining? In brief, the generic term data mining refers to extracting or "mining" knowledge from large amounts of data. According to Suchita and Rajeswari (2013), it is a process of analyzing data from different perspectives and summarizing it into important information in order to identify hidden patterns from a large data set. The main functions of data mining are applying various techniques and algorithms in order to detect and extract patterns of a given stored data set (Jiawei et al., 2012). In previous studies, there have been several studies that present a comprehensive review of the data mining applications. For instances, see Florin, G. (2011) and Jiawei et al. (2012). According to Barros and Verdejo (2000) and Jiawei et al. (2012), data mining can be classified into various algorithms and techniques, such as classification, clustering, regression, association rules, etc, which are used for knowledge discovery from databases. The data mining algorithms must be, theoretically, fully realized and well described with regard to educational data analysis and must be proved prior these algorithms used practically. In the next section, a brief overview of some of these techniques is given. For more detailed discussion of data mining, see Michael and Gordon (2004) as well as Ian and Eibe (2005). Here, we have to make it clear that, this research restricts attention to only consider the classification task for assessing student's performance, specifically the C4.5 algorithm is the main focus of this paper.

In this paper, student's information, such as their degrees in the previous academic records (annually) are collected to predict the performance at the end of the last year based on various attributes. The study was done on the data set that has 124 graduate students. Further, we have identified the important and necessary attributes that impact the student's academic performance. An application study is implemented using the WEKA software and real time data set available in the college premises. The paper aims to predict the student's performance in the faculty result based on the basis of his/her performance throughout the study period. The paper is organized as follows: In Section 2, a background for data mining is provided, followed by a particular focus on the decision trees modeling based on the C4.5 algorithm. Section 3 consists of our application schemes study, including a description of the data set used in the analysis. The findings are next interpreted and discussed in Section 4. The study concludes in Section 5, with a brief description of some concluding remarks.

2. DATA MINING ALGORITHMS

As stated in Ian and Eibe (2005), data mining algorithms have become a huge technology system after years of development. Generally, data mining has the following basic topics: (1) Classes: stored data are used to locate objects in predetermined groups; (2) Clusters: data items are grouped according to logical relationships or consumer preferences; (3) Associations analysis: data can be mined to identify associations; (4) Sequential patterns: in this topic data sets would be mined to anticipate behavior patterns and trends; and (5) Prediction: in this topic, data can be fitted in order to have their trends and behavior as well as to estimate the future behavior depending upon the historical data sets stored in data warehouse. As discussed earlier, the classification task is a focus of this article. Sun et al. (2008) stated that classification is a systematic technique based upon the input data to establish a classification model. Moreover, the classification examples consist of the following algorithms: decision tree, rule-based, Naive bayes, etc. However, despite these number of classification methods, we focus on the C4.5 algorithm which is one of the decision trees classification algorithms as it has been the data mining approach of choice. In fact, the classification task aims to construct a model in training
data set in order to estimate the class of future objects that have unknown class label. There are two broad topics in classification; these are: (1) Preparing the data for classification and prediction; and (2) Comparing classification and prediction methods (Ian and Eibe, 2005). Furthermore, classification employs a set of pre-classified examples to develop a model that can classify the population of records at large. In general, the data classification process involves learning and classification. In learning the training data are analyzed by classification algorithm. In addition, the classification test data can be applied to estimate the accuracy of the classification rules (Florin, 2011).

2.1. Decision Trees Modeling - The C4.5 Algorithm

Decision tree modeling is one of the classifying and predicting data mining techniques, belonging to inductive learning and supervised knowledge mining. It is a tree-diagram-based method, depending on two manners; the node on the top of its tree structure is a root node, and nodes in the bottom are leaf nodes. Each leaf node would be having a target class attribute. There would be a path node, for each every leaf node, of multiple internal nodes that have attributes based on a root node. Further, the considered path creates some rule requested for determining the classification of unknown data set. Moreover, most of decision tree algorithms contain two-stage task, i.e., tree building and tree pruning. In terms of the tree constructing stage, a decision tree algorithm can use its unique route in order to specify the valuable attribute, so as to split training data set.

Finally, the last position regarding this stage would be that data that included in the split training subset belong to only one specific target class. Recursion and repetition upon attribute selecting and set splitting will fulfill the construction of decision tree root node and internal nodes. On the other hand, there are some principal data in training data set can give an improper branch on decision tree building; this is usually denoted by the term of over-fitting. Therefore, after building a decision tree, it has to be pruned to remove improper branches, so as to enhance decision tree model accuracy in predicting new data. Among developed decision tree algorithms, the commonly used ones include ID3, C4.5, CART and CHAID. The C4.5 algorithm is an extension of the ID3 (Iterative Dichotomiser 3, it is a simple decision tree learning algorithm developed by Quinlan (1986)) algorithm, it uses information theory and inductive learning method to construct decision tree. C4.5 improves ID3, which cannot process continuous numeric problem. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. Further details of these algorithms can be found in Kass, G. V. (1980), Ian and Eibe (2005) and Sun et al. (2008). Decision trees based on the C4.5 algorithm is a commonly used classification techniques which extract relevant relationship in the data. Overall, the C4.5 algorithm is refers to a program that generates a decision tree depending upon a set of labeled input data. Further, the decision trees modeling created by this algorithm can be used for classification, and for this reason, the C4.5 algorithm is often defined as a statistical classifier. The C4.5 algorithm makes decision trees using a set of training data, taking into account the concept of information entropy. The training data can be defined as a set \( S = s_1, s_2, \ldots \) of already classified samples. Thereafter, each sample \( S_i = x_1, x_2, \ldots \) is a vector, where \( x_1, x_2, \ldots \) denotes attributes of the sample. Then the training data is augmented with a vector \( C = c_1, c_2, \ldots \) denotes the class to which each sample belongs.
3. APPLICATION STUDY

3.1. Data Description

In this paper, we consider student's data set that is pursuing Bachelor of Statistics, Actuarial Science degree from Faculty of Mathematical Sciences and Statistics in Alneelain University. The variables used for assessing the student's performance as well as for building a predicted models in the faculty were degree1, degree2, degree3, degree4 and degree5, corresponding to the students degrees in the period from 2008 to 2013. The number of graduates selected was 124. As discussed earlier, the study was focused on the previous academic records of the students. The first fourth Student degrees have been used in order to predict the student degrees with respect to the fifth degree. Of these 124 records in our data set, we set up that each record has five numerical attributes. These attributes provide the annual degrees in which values ranged from 1 to 124. The data has been preprocessing in three stages:

1. Convert the first fourth degrees into nominal data type according to the following syntax:

   if in [40..59]: Pass,
   if in [60..69]: Good,
   if in [70..79]: V.Good
   otherwise: Excellent

2. Handling the missing attribute information using the imputation technique. Discussions with respect to the technical details of the imputation technique are given by Satty and Mwambi (2012).

3. Divide the class label into the three broad classes. This has been done using the following syntax:

   if less than 59: class C (students need extreme improvement to their degrees)
   if in [60..79]: class B (students need a little bit improvement in their performance)
   otherwise: class A (this class includes those students who were doing well)

3.2. The Weka Software

According to Remco et al. (2012), WEKA is defined as an open source application that is freely available under the GNU general public license agreement. Firstly, this software has been originally written in C, the WEKA application thereafter has been completely re-written in Java, and is compatible with almost every computing platform. Generally speaking, WEKA can be defined as a computer program that has been developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data sets gathered from agricultural fields. It can be used to apply many different data mining tasks such as data preprocessing,
classification, clustering, and so on. However, in this paper, we only placed a particular emphasis on considering the C4.5 algorithm as it is a commonly used classification algorithm. More details of WEKA, including its characteristic system, file format, system interface, the mining process can be found in Remco et al. (2012). This software deals with the data sets that have specific formats, such as the so-called ARFF (Attribute-Relation File Format (ARFF), CSV (Comma Separated Values) and C4.5.s format, as few examples. These specific formats have been taken into account in dealing with this paper.

3.3. Fitting A C4.5 Algorithm

The main objective of implementing the C4.5 algorithm is to give a model that can be used for estimating the class of the unknown tuples as well as records. To do so, we used the following steps that represent basic principle of working for this classifier: (1) We give a training set which contains the training results together with their linked class label; (2) Therefore, we construct the classification model by carrying out the learning algorithm that can be used in respective technique; and (3) The model built is carried out using the test set that contains of the tuples not having a link with class label. This algorithm has been carried out using the CRISP process. CRISP refers to CRoss Industry Standard Process, which contains six stages. Figure 2 displays the link between them.

![Figure 1: CRISP Process](image-url)
For a binary decision problem, a classifier labels examples as either positive or negative. The confusion matrix or can be used constructed to make the decision that can be made by classifier. This matrix consists of four categories: True positives (TP): these are instances correctly labeled as positives; false positives (FP) correspond to negative instances incorrectly labeled as positive; true negatives (TN) refer to negative correctly labeled as negative; and false negatives (FN) correspond to positive instances incorrectly labeled as negative. We further define TPR as the true positives rate, which is equivalent to Recall (would be briefly visited below). This matrix builds the so-called recall and precision measures. Recall can be computed as Recall = TP/(TP+FN). Precision measures that fraction of examples classified as positive that are truly positive. It can be calculated as Precision = TP/(TP+FP). In fact, there are several basic measures that can be used to assess the student's performance. Such measures are readily usable for the evaluation of any binary classifier. Consequently, to assess the performance of the data set mentioned above, we use these evaluation criterions depending on the next measures: (1) Accuracy: it is defined as the number of correct predictions divided by the total number of predictions; and (2) Error rate: It refers to the number of wrong predictions divided by the total number of predictions. Furthermore, the description of each measure here is shown below: The correctly classified instances show degree of test instances that were correctly classified (Accuracy). The incorrectly classified instances show age of test instances that were incorrectly classified (Error Rate); (3) The Kappa statistic: It was introduced by Cohen (1960). Bartko and Carpenter (1976) has stated that the Kappa statistic refers to a normalized statistic measure of agreement. This measure of agreement can be computed by dividing two quantities; the first quantity is a agreement expected by chance away from the observed agreement between the classifier and actual truth and the second quantity is the maximum possible agreement. The possible value for Kappa lies in the range [-1, 1] although this statistics usually falls between 0 and 1. The value of 1 indicates perfect agreement, however, the value of 0 indicates that the agreement no better than expected by chance. Therapy, when $K$ has a value greater than a value of 0, it implies that the classifier is doing better compared to chance, and therefore indicating perfect agreement at $K = 1$; otherwise, if the value of $K$ is 0, then it denotes the chance agreement. A kappa statistic associated with the negative rating gives worse agreement than that expected by chance. Now, let $P_a$ and $P_e$ denote the percentage agreement and expected chance (hypothetical) agreement, respectively. Consequently, this statistic can be expressed as follows:

$$K = (P_a - P_e) / (1 - P_a).$$

In our analysis, for computing k, we have the total instances = 124. (4) F-Measure combines recall and precision scores into a single measure of performance. It can be computed as $F$-Measure=$2\times$(recall$\times$precision) / (recall + precision). (5) ROC area (Receiver Operator Characteristic) is commonly used to provide findings for binary decision problems in data mining. Using it together with the recall and precision measures we can get a more informative picture of the C4.5 algorithm.
4. RESULTS AND DISCUSSION

From the C4.5 classification algorithm, the decision tree is constructed, depending on the most effective attribute(s) is/are given using the so-called Entropy and the Gain information. Hence, to achieve this construction, we needed to compute the entropy for each feature depending upon the training images given by the C4.5 algorithm and measure the information gained of every feature and finally take maximum of them in order to be considered as a root (Andreas and Zantinge, 1996). Entropy\( (S) = \sum_{i=1}^{n} -P_i \log_2 P_i \), and the Information Gain is given by

\[
Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_v|}{|S|} Entropy(S_v),
\]

where \( P_i \) is the probability of a system being in cell \( i \) of its phase space, \( \sum_{i=1}^{n} -P_i \log_2 P_i \), gives the entropy of the set of probabilities \( P_1, P_2, \ldots, P_n \), \( \sum \) is over each value \( v \) of all the possible values of the attribute \( A \), \( v \) is the subset of \( S \) for which attribute \( A \) has value \( v \), \( |S_v| \) is the number of element in \( S_v \), \( |S| \) is the number of element in \( S \).

Figure 2: Decision tree for students degrees data set
Anyhow this paper is interested in finding out the relationships between the considered degrees attributes. Therefore, decision tree is displayed in Figure 1. This figure shows that, depending upon the information gain, the attribute deg3 is having the maximum. Therefore, this degree has located at the top of the decision tree (decision tree algorithms use the gain value to start splitting the tree with attribute having high gain and so on). The number of leafs from the decision tree output was 16. The tree size that is obtained was 21; the time taken to build our model was 0.02 seconds. Moreover, the figure is showing that partial tree is the result of fitting the C4.5 classification algorithm model, in which the tree consists of 5 leaves marked with $L_1, L_2, L_3, L_4$ and $L_5$.

1. Deg3 = pass
2. Deg4 = Good: C (12.0/4.0)
3. Deg4 = pass
4. Deg1 = pass: C (8.0/2.0)
5. Deg1 = Good: B (6.0/2.0)
6. Deg4 = V.Good: B (2.0)
7. Deg3 = Good: B (72.0/8.0)

The leaf $L_1$ contains instances (12, 4) in the row number 2, node Deg4. Therefore, in this leaf there were 16 records from the data set have classified in class C. The leaf $L_2$ contains instances (8, 2), which is to say that 10 records were classified in class C. $L_3$ contains instances (6, 2) in row number 5, Deg1 = Good, this implies that 8 records have been classified in class B. $L_4$ consists of instances (2) in row number 6, node Deg4 = V.Good, which means that this leaf has 2 records that have classified in class B. Finally, $L_5$ consists of instances (72, 8) in row number 7, node Deg3 = Good, which means that 80 records were classified in class B. The results further yielded the confusion matrix. From this matrix we extract the following findings: (1) 25 records were classified in class C, thus 20% belong to C, 14 of these records are TP with rate of 56%; (2) 91 records have been categorized to be in class B. This refers to that 73% belong to B, and 77 records were TP with the rate of 85%; and (3) 8 records have been seen in class A, meaning 6% belong to this class. 5 records of them were TP under the rate of 63%.

The results for computing the accuracy and error rate measures are displayed in Table 1. By looking at this table, we find that (as the number of instances was equaled to 106) the accuracy = (106/124)*100 = 85.4839%, and the number of incorrectly classified equals 18, the error rate therefore = (18/124)*100 = 14.5161%. As we see in Table 1, in order to fit the C4.5 algorithm, we provide the training set to build a predictive model. This training set consists of the predictor attributes as well as the prediction (class label) attribute. First, we use the training set in the preprocess panel, followed by the selection of the C4.5 algorithm. Thereafter, we selected the so-called the 10 fold cross validation choice. Second, we apply the same procedure on our testing set to check what it predicts on the unseen data. For that, we select "supplied test set" and choose the testing data set that we created. Finally, we run the C4.5 again and we notice the differences in accuracy. Note that when the instances are used as test data, the correctly/incorrectly classified
instances can determine the case. Depending on these findings, we see that 85.4839% can be considered as a good percentage to achieve the main goal of this paper. Turning to the error rates displayed in the table, we see that the error rates are the same for both training and supplied tests. This indicates that the considered algorithm was doing well for both tests. However, for cross validation folds as well as for percentage split 66%, the error rates were different, which is to say that C4.5 was effective with respect to cross validation fold as it has a lower error. This can be justified by the fact that an algorithm will be preferred when it has a lower error rate, namely it has more powerful classification capability and ability in terms of student's performance. On the other hand, the Kappa statistic that we obtained was 0.6327, which is to say that the used algorithm in our model is well doing as the Kappa statistics is greater than 0 (see, Cohen, 1960 in terms of interpreting a Kappa statistic). The results of recall, precision, F-measure and ROC area are displayed in Table 2. The findings yield that the recall and precision present estimates closer to each other. Note that in precision and recall measures, since there is a variation in the level of recall measurement, the precision measurement cannot be linearly changed. This can be justified by the fact that the fact that there is a replacement concerning FP and FN in the denominator of the precision metric. As we know higher precision as well as F-measure are better. Thus, as given in Table 2, the findings were high (above 70%) leading to that fact that the C4.5 algorithm is an effective and reliable technique to be recommended.

Table 1: Testing options

<table>
<thead>
<tr>
<th>Training option</th>
<th>Correct classify instance %</th>
<th>Incorrect classify instance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>85.4839%</td>
<td>14.5161%</td>
</tr>
<tr>
<td>Supplied test</td>
<td>85.4839%</td>
<td>14.5161%</td>
</tr>
<tr>
<td>Cross validation folds</td>
<td>77.4194%</td>
<td>22.5806%</td>
</tr>
<tr>
<td>Percentage split 66%</td>
<td>76.1905%</td>
<td>23.8095%</td>
</tr>
<tr>
<td>Kapa = 0.6327</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Detailed accuracy for each class–classification using the C4.5 algorithm

<table>
<thead>
<tr>
<th>TP rate</th>
<th>FP rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>ROC area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.56</td>
<td>0.061</td>
<td>0.7</td>
<td>0.56</td>
<td>0.622</td>
<td>0.832</td>
<td>C</td>
</tr>
<tr>
<td>0.934</td>
<td>0.333</td>
<td>0.885</td>
<td>0.934</td>
<td>0.909</td>
<td>0.844</td>
<td>B</td>
</tr>
<tr>
<td>0.875</td>
<td>0.009</td>
<td>0.875</td>
<td>0.875</td>
<td>0.875</td>
<td>0.992</td>
<td>A</td>
</tr>
<tr>
<td>Weighted avg</td>
<td>0.855</td>
<td>0.257</td>
<td>0.847</td>
<td>0.855</td>
<td>0.849</td>
<td>0.851</td>
</tr>
</tbody>
</table>
5. CONCLUSION

In this study, we have placed a particular emphasis on the so called data mining algorithms, but focus the bulk of attention on the C4.5 algorithm. Our goal was to build a predicted model that can be used to improve the student's academic performance. In order to achieve this goal, data from the previous students' academic records in the faculty have been used to illustrate the considered algorithm in order to build our predicted model. In spite of the fact that there are several other classification algorithms in the literature, the C4.5 approach was the common data mining technique of choice for the primary analysis for dealing with students performance prediction because of its simplicity as well as the ease with which it can be implemented. Here we refer to statistical software such as, SPSS and SAS. Thus, the C4.5 approach might become attractive in specific circumstances. We here believe that the C4.5 algorithm can be recommended as a default tool for mining analysis. The findings in general revealed that it is possible to predict the probability of getting a degree within the estimated period according the degree of a graduate in the attributes performance. In conclusion we submit that the algorithm described here can be very helpful and efficient if there is an application study regarding the assessment of students' performance, where both kind of knowledge is required (association among attributes and classification of objects).

REFERENCES