Predicting Instructor Performance in Educational Institution using Data Mining Techniques

Dr. P. Tamjue Selvy¹, D. Madhumathi², B. Preetha³, S. Sarika⁴
Associate professor¹, UG Scholar²,³,⁴
Computer Science and Engineering
Sri Krishna College of Technology, India

Abstract:
Data Mining applications focuses on collecting knowledge from the databases or data warehouses. The Knowledge is nothing but the particular information what we need. Data Mining applications play a major role in solving educational problems in higher education. Nowadays, Educational data mining is an emerging discipline concerned with different approaches includes predicting instructor performance and improve the quality of education process to enhance students performance. One of the common tool to evaluate the instructors’ performance in a course is through surveying students’ responses about the course and its instructor through a questionnaire. Four different Classification techniques include - (i) Decision tree algorithm (ii) Support Vector machine (iii) Artificial Neural Networks and (iv) Discriminant Analysis. They are used to build classifier models. Naive Bayesian classifier is the proposed classification technique is best with respect to accuracy, precision, recall and specificity. Alumni reference, staff reference and present student feedback are the main factors in predicting instructor performance. Overall performances are compared over a data set. The proposed method is best with respect to accuracy, precision, recall, and specificity when compared to previous techniques and also it reduces the average error rate in classification.

Keywords: Data mining, Classifier model, instructor prediction, Discriminant analysis.

1. INTRODUCTION

1.1 Data Mining
Data mining is the process of finding interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories [1]. The term data mining is also named as Knowledge Discovery. Analysis and prediction with the help of data mining techniques have shown noteworthy results in many areas. It can be very effective in Education System as well. It is a very powerful tool to reveal hidden patterns and precious knowledge.

1.2 Educational Data
Educational institutions are becoming more competitive because of the rapid growth of many institutions. To be competitive, the institutions should identify their own potentials hidden and implement a technique to bring it out. The higher education institutions has potential knowledge such as academic performance of students, potential knowledge of the faculty and many other information in a hidden form[1]. The hidden knowledge is extracted using technique called Knowledge Discovery process. To improve the learning capabilities of the students, the instructors should be capable of monitoring the overall performance of each student, separately and dynamically adjust their teaching methodologies on students and to take immediate decisions to improve learning outcome of students.

1.3 Classification
Classification is a two step process consisting of training and testing. In the first step, a model is constructed by analyzing the data tuples from training data having a collection of attributes. For every tuple in the training data, the significance of class label attribute is understood. Classification rule is applied on training data to form the model. In the next step of classification, test data is in use to examine the accuracy of the model. If the accuracy of the model is suitable then the model can be used to classify the unknown data tuples. The fundamental techniques used for classification are decision tree classifier, neural networks, rule based classifier and Lazy based classifier.

2. RELATED WORK

In this section, we discuss existing solutions for the instructor performance predictor and several methods are reviewed.

2.1 Knowledge Discovery
General reviews of the Educational Data Mining (EDM) research field are provided by Romero and Ventura [6]. In their paper, they examined the application of data mining to traditional educational systems, particular web-based courses, adaptive and intelligent web-based educational systems. Another review was carried out by Baker and Yacef[7] covering the period between 2005 and 2009. They reviewed the current trends in the field of EDM, considered the procedural profile of research in the early years of EDM, compared to 2008 and 2009, and discussed trends and modifications in the research conducted by this community. Another more recent survey was done by Raji and Malaya [1], which focused on components, research trends (1998 to 2012) of EDM highlighting its associated Tools, Techniques and educational Outcomes. They also highlighted the Challenges in EDM. A study by Varun and Arupama[9] observed the application of data mining techniques in higher educational institution to extract useful information from enormous data sets and provided analytical tool to view and use this information for decision making processes by taking real life examples.
2.2 CHAID and CART algorithm

Mardikyan and Badur [13] conducted a study to investigate the factors associated with the assessment of instructors teaching performance using two different data mining techniques: decision trees and regression analysis. CHAID and CART algorithms were applied for decision trees and the stepwise regression method was used for regression analysis. As a result, they found that instructors help the student outside the lectures, grade exams fairly and on time receive higher evaluations.

2.3 Naïve Bayesian algorithm

Chin-Chia Hsu and Tao Huang [14] conducted a study on the use of data mining technology to evaluate student’s academic achievement via multiple channels of enrollment like joint athletic enrollment, recruitment enrollment and application enrollment. A similar study was carried out by Ososifan and Olamiti [3] where they explored the academic background in relationship with the students performance in a Nigerian university. Their study illustrated that the grade obtained from senior students examination in mathematics is the highest determinant of students’ performance using the C4.5 learning algorithm in building the model of the student’s performance.

2.4 Motivation

In previous section, different classification techniques are used. The results of each classification technique varies. These methods either lacks in accuracy or irreliability. To rectify these disadvantages, the proposed scheme uses Naïve Bayesian classifier for more accuracy. It reduces the average error rate to maximum extent.

3. PROPOSED SYSTEM

In the proposed system, Bayesian classification approach is used to predict the instructor performance. The Proposed Naïve Bayesian classifier is best with respect to accuracy, precision, recall and specificity. The system is designed by collecting datasets of the instructors performance from the senior students of the institution. Figure 1 shows the overall process of the system.

![Overall Process of the System](http://ijesc.org/)

Figure 1. Overall Process of the System

Thereby grouping instructor performance based on certain condition as best, good, average and poor. This system provides an efficient analysis on instructor performance by data collection and result prediction. Naïve Bayesian technique is very accurate than many other classification techniques. Advancements in the Internet technologies have introduced various learning methods[12]for the students to gather more knowledge in a shared and collective way, which has changed the way of gaining knowledge.

3.1 Attribute Selection using Decision Tree

Decision tree is a common method used in data mining. It is an efficient method for producing classifiers from data. It is a tree-structured method consists of a set of attributes to test in order to predict the output. It is used for determining the optimum course of action, in situations having several possible alternatives with uncertain outcomes[5]. There are two rules implemented in the decision tree, the first rule says that, if the trainees have implemented experience they have gained in their classrooms and give the trainees possibility of the use of the information gained from this training in their teaching in the future and begins the session with knowledge of (past experiences) of the trainees and the course length is sufficient for the benefit of the trainee as required, the session has improved the professional competence can be predicted as ok by a high proportion[9]. The second rule is same as the first rule, but the session has improved the professional competence that can be predicted as ok by a medium proportion.

3.2 Classification Using SVM

Support vector machine is a supervised learning method mainly used for classification. It classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The largest margin between the two classes is the best hyperplane for an SVM[7]. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The support vectors are the data points that are closest to the separating hyperplane. In a straight line case, a simple equation [1 & 2] gives the formula for the maximum margin hyperplane as a sum over the support vectors.

\[ x = b + Nbr of Support Vectors \times 1 \]
\[ \sum = 1 \text{.} y_{i} \text{.} a(i) \cdot a = b \sum i = 1 \text{.} y_{i} \text{.} a(i) \cdot a \]

Where \( a \) is the Lagrange multiplier and \( b \) is the parameter. \( x \) and \( y \) are the points in the graph. These are kind of a vector product with each of the support vectors, and the sum. It's pretty simple to calculate this maximum margin hyperplane once you've got the support vectors.. It depends on the support vectors. None of the other points play any part in this calculation. Support Vector Machine method has attained the highest prediction accuracy in recognizing students at risk of failing[15].

3.3 Estimating Probabilities Naïve Bayesian

Naïve Bayesian classifier is a technique for estimating probabilities of individual variable values, given a class, from training data and then allow the use of these probabilities to classify new entities[11]. The session has improved the professional competence that resulted from applying the Naïve Bayesian Kernel classifier.

4. EXPERIMENTAL RESULTS

4.1 Instructor Dataset

As a part of preprocessing and dataset preparation the Date Of Work attribute in the data set contains a large number of
values. So we grouped it into Four categorical segments - New teacher, Teacher, First teacher and Expert teacher as follows:

IF Date of work<=1 year He's "New teacher" IF Date of work<=10 &>1 years he's "Teacher "IF Date of work<=15 &>10 years he's "First teacher "IF Date of work >15 years he's "Expert teacher".

4.2 Instructor Feedback Report

The Performance of the proposed system is demonstrated using visual studio 2010 and MS SQL server database. Synthetic datasets are used for predicting the instructor performance. A set of questionnaire is prepared and those are filled by students and it is been evaluated. By evaluating the feedback the teaching methodology is improvised as a result, the quality of education can be improved.

4.3 Performance Analysis

In this section, we will discuss about the performance analysis in terms of accuracy, precision, recall and specificity. The class variables values may be assumed as

P→Positive  
N→Negative  
TP→ True Positive  
TN→ True Negative

Table 1 depicts the calculated values of accuracy, precision, recall and specificity of the classification models.

**Accuracy**: Accuracy values, which assess the effectiveness of the models, are all at least approximately 90%. The equation[3] is used to calculate the accuracy.

\[ \text{Accuracy} = \frac{TP + TN}{P + N} \rightarrow [3] \]

**Precision**: Precision, which assesses the predictive power, again indicated C5.0 as the best classifier; however, SVM also show high predictive power. The equation [4] is used to calculate precision.

\[ \text{Precision} = \frac{TP}{TP + FP} \rightarrow [4] \]

**Recall**: Recall values which indicate the sensitivity and true positive rate (TPR) of the models, differ among classifiers. It measures the rate of positives predicted as positives by the classifier.

<table>
<thead>
<tr>
<th>Performance metrics</th>
<th>Classifiers</th>
<th>C5.0</th>
<th>SVM</th>
<th>Naive Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td></td>
<td>91.5</td>
<td>90.2</td>
<td>93.5</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>91</td>
<td>91.5</td>
<td>92.5</td>
</tr>
<tr>
<td>recall</td>
<td></td>
<td>89</td>
<td>90</td>
<td>92</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>90.5</td>
<td>87</td>
<td>91</td>
</tr>
</tbody>
</table>

![Figure 3. Comparison of Precision, Accuracy, Recall and Specificity](http://ijesc.org/)

Figure 3. Comparison of Precision, Accuracy, Recall and Specificity
The equation [5] is used to calculate recall.

\[
\text{Recall} = \frac{TP}{P} \rightarrow [5]
\]

**Specificity:** Specificity measures the rate of negatives predicted as negative by the classifier. The figure 3 shows the comparison of precision, accuracy, recall and specificity over all the methods used. The equation [6] is used to calculate specificity.

\[
\text{Specificity} = \frac{TN}{N} \rightarrow [6]
\]

5. CONCLUSION AND FUTURE WORK

The proposed system provides the accurate performance of the staff. This system provides the comparison of the performance of the staff with the dataset. So as to generate the effective instructor performance predictor. Optimized classification techniques can be used to enhance the classification accuracy. Thus through these methods less false positive rate and improved prediction accuracy in diagnosing has been achieved. This work can be extended to further improved predictor system using optimization techniques.

6. REFERENCES:


7. Short Bio Data for the Author

Prof. P.TamijSelvy received B.Tech (CSE), M.Tech (CSE) in 1996 and 1998 respectively from Pondicherry University. She completed her Ph.D under Anna University, Chennai in the year 2013. Since 1999, she has been working as faculty in reputed Engineering Colleges. At present, she is working as Associate Professor in the department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore. Her Research interests include Image Processing, Data Mining and Pattern Recognition

Ms.D.Madhumathi is currently pursuing Bachelor of Engineering degree in Computer Science and Engineering under Anna University, Coimbatore, India. Her area of interest is Data Mining.
Ms. B. Preetha is currently pursuing Bachelor of Engineering degree in Computer Science and Engineering under Anna University, Coimbatore, India. Her area of interest is Data Mining.

Ms. S. Sarika is currently pursuing Bachelor of Engineering degree in Computer Science and Engineering under Anna University, Coimbatore, India. Her area of interest is Data Mining.