Work in Progress - A Decision Tree Approach to Predicting Student Performance in A High-Enrollment, High-Impact, and Core Engineering Course

Ning Fang and Jingui Lu
nfang@engineering.usu.edu, lujg@njut.edu.cn

Abstract - **This paper aims at developing a decision tree model to predict student performance in engineering dynamics – a high-enrollment, high-impact, and core engineering course. This study is innovative because no prior literature exists on the same topic. Three research contributions are made: 1) Nine "if-then" decision rules were generated to predict student performance in engineering dynamics. 2) It is revealed that a student's score in Statics and cumulative GPA play a significant role in governing student performance in engineering dynamics. 3) It is revealed that the decision tree predictions are more accurate than the predictions from the traditional multivariate linear regression technique.**

Index Terms - Decision tree, Engineering dynamics, Predictive modeling, Student performance.

INTRODUCTION

Engineering dynamics is a high-enrollment, high-impact, and core engineering course that almost every mechanical or civil engineering student is required to take. This course is widely regarded by students as one of the most difficult courses, and many students fail because it covers a broad spectrum of foundational engineering concepts and principles, such as motion, force and acceleration, work and energy, impulse and momentum, and vibrations for both a particle and a rigid body. Prediction of student performance in dynamics is important because not only does it help the instructor develop effective course curriculum and teaching strategies, but facilitates students' increased understanding and development of effective learning strategies [1].

A decision tree is a flow-chart-like tree structure [2]. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of "if-then" rules (rather than abstract mathematical equations), making the results easy to interpret. The decision tree has received growing attention in engineering, business, medicine, and education research.

The present study aims at developing a decision tree model to predict student performance in engineering dynamics. The present study is innovative because no prior literature exists that applies the decision tree approach (or any other data mining and statistical approaches) to model student performance in engineering dynamics. In prior literature, a decision tree was used on business management, production planning and control, and student enrollment analysis, etc. [1, 3], but not on engineering dynamics.

The decision tree model developed in the present study includes one outcome/dependent variable (i.e., a student's score in dynamics) and five predictor/independent variables including the student's cumulative GPA and scores in four prerequisite courses: Engineering Statics, Calculus I, Calculus II, and Physics. The five predictor variables measure students' cognitive levels and problem-solving skills. Students' non-cognitive factors (such as learning style, motivation and interest, and time devoted to learning), the instructor's teaching effectiveness, as well as teaching and learning environment will be included in the future work beyond this exploratory, work-in-progress study.

RESEARCH QUESTIONS

The three research questions of the present study are:

1. What specific "if-then" rules can be generated to predict student performance in engineering dynamics?
2. To what extent a student's cumulative GPA and performance in prerequisite courses affect student performance in engineering dynamics?
3. Are decision tree predictions more accurate than the predictions from traditional statistical approaches?

DATA COLLECTIONS

A validated total of 750 data records were collected from 125 students in Semesters A (45 students) and B (80 students), including 113 males and 12 females. Student majors were: mechanical engineering (61), civil engineering (40), and bioengineering and other majors (24).

The collected data included student cumulative GPA (numerical values 0.0-4.0) and scores (letter grades A, A-, B+, B, B-, C+, C, C-, D+, D, or F) in Dynamics and four prerequisite courses: Statics, Calculus I, Calculus II, and Physics. Research has shown a strong co-relationship between student prior achievements and Dynamics course performance. Each student was associated with six data records. For 125 students, there are 6×125=750 data records.

DECISION TREE MODELING AND VALIDATION

Student records were divided into the training dataset and the validation dataset. In Semester A, the training dataset contained the records from the first 15 students (in alphabetical order of their last names in the class), and the records from the remaining 30 students were used as the validation dataset. In Semester B, the training dataset had the records from the last 16 students (in alphabetical order of their last names in the class), and the records from the remaining 64 students were used as the validation dataset.

A decision tree algorithm [2] was employed to construct a decision tree for each semester. The splitting criterion was entropy reduction. The maximum depth of the tree was five. The maximum number of branches from a node was three. Because the decision tree algorithm only deals with categorical (ordinal or nominal) data, the student cumulative GPA (numerical values 0.0-4.0) was converted into letter grades (A, B, C, DF). To reduce the number of decision rules so that they can be easily employed, Grade A included A and A-; Grade B included B+, B, and B-; Grade C included C+, C, and C-; Grade DF included D+, D, and F.

Nine decision rules were generated for each semester. Take Semester A as an example. Some decision rules generated from the training dataset are given below, where PRP stands for the prediction reliability percentage:

Rule I: IF Statics: B; and Calculus II: B
THEN Dynamics: B (PRP: 100%)

Rule II: IF Statics: B; Calculus II: C; and Physics: B
THEN Dynamics: C (PRP: 66.7%)

Rule III: IF Statics: C; and GPA: B
THEN Dynamics: B (PRP: 91.7%)

Rule IV: IF Statics: DF
THEN Dynamics: DF (PRP: 100%)

The relative importance level of each predictor variable is in the following order (for Semester A): Statics: 100%; Cumulative GPA: 74.5%; Physics: 51.7%; Calculus II: 43.2%; and Calculus I: 43.2%. Thus, a student's score in Statics and cumulative GPA are the two most important variables that govern student performance in dynamics.

To validate the developed decision tree models, the validation dataset for each semester was employed. The results show that for Semester A, the decision tree model accurately predicts the Dynamics score of 25 students (83.3%) out of 30 students. For Semester B, the decision tree model accurately predicts the Dynamics score of 55 students (85.9%) out of 64 students.

COMPARISON OF THE DECISION TREE APPROACH WITH A TRADITIONAL MULTIVARIATE STATISTICAL APPROACH

Because the outcome variable (i.e., a student's score in engineering dynamics) has four levels of variations (A, B, C, and DF), the logistical regression technique that only deals with binary categorical data no longer applies in the present study. Thus, the multivariate linear regression technique was employed for comparison purposes. All letter grades (A, B, C, and DF) were converted to their numerical values: A=4.0, B=3.0, C=2.0, and DF=1.0. The cumulative GPA took its actual numerical values (0.0-4.0), such as 1.55 and 2.78.

The same training dataset was employed to establish the linear regression model for each semester. Using Semester A as an example again, the linear regression model is:

$$\text{Dynamics score} = -2.039 + 1.654\times\text{GPA} - 0.253\times\text{Statics score} - 0.500\times\text{Calculus I score} + 0.239\times\text{Calculus II score} + 0.329\times\text{Physics score} \quad (1)$$

The coefficient of determination of (1) is 0.781. The negative coefficients of two predictor variables (Statics and Calculus I) imply that a student earns a higher score of Dynamics if their scores in two prerequisite courses are low, making no real-world sense. The same validation dataset was employed to test the prediction accuracy of (1). Table I shows the comparison of prediction accuracy.

TABLE I
COMPARISON OF PREDICTION ACCURACY

Semester	Decision Tree	Linear Regression
A	83.3%	66.7%
B	85.9%	71.9%

Table I illustrates that, compared to the linear regression approach, the decision tree approach improves the prediction accuracy by 16.6% in Semester A and 14% in Semester B.

CONTRIBUTIONS AND CONCLUSIONS

The decision tree approach is employed in this exploratory study to predict student performance in Dynamics. The answers to the three research questions are:

1) A set of specific "if-then" rules were generated to predict student performance in engineering dynamics.

2) A student's score in Statics and cumulative GPA play the two most significant role in governing student performance in dynamics.

3) The decision tree predictions are more accurate than the predictions from the linear regression technique.

REFERENCES

[1] Nghe, Nguyen T., Paul Janecek, and Peter Haddawy. "A Comparative Analysis of Techniques for Predicting Academic Performance." In Milwaukee, WI. 10-13 October 2007. *Proceedings of the 2007 Frontiers in Education Conferences*, p. T2G-7.

[2] Han, Jiawei and Micheline Kamber. 2007. *Data Mining: Concepts and Techniques*. San Fransisco, CA: Morgan Kaufmann publishers.

[3] Mendez, Guillermo, Trent D. Buskirk, Sharon Lohr, and Susan Haag January 2008. "Factors Associated With Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests." *Journal of Engineering Education*. Vol. 97, pp. 57-70.

AUTHOR INFORMATION

Ning Fang Department of Engineering and Technology Education, College of Engineering, Utah State University, Logan, Utah 84322 nfang@engineering.usu.edu

Jingui Lu CAD Center, Nanjing University of Technology, Nanjing 210009, Jiangsu Province, P. R. China, lujg@njut.edu.cn