2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems

Predicting Student Academic Performance

Kin Fun Li, David Rusk and Fred Song
Electrical and Computer Engineering
University of Victoria
Victoria, Canada
kinli@uvic.ca

Abstract— **Engineering schools worldwide have a relatively high attrition rate. Typically, about 35% of the first-year students in various engineering programs do not make it to the second year. Of the remaining students, quite often they drop out or fail in their second or third year of studies. The purpose of this investigation is to identify the factors that serve as good indicators of whether a student will drop out or fail the program. In order to establish early warning indicators, principal component analysis is used to analyze, in the first instance, first-year engineering student academic records. These performance predictors, if identified, can then be used effectively to formulate corrective action plans to improve the attrition rate.**

Keywords: performance prediction; student attrition; student academic performance; engineering education; principal component analysis

I. ENGINEERING STUDENTS PERFORMANCE

Engineering programs worldwide traditionally have a relatively high attrition rate. There is no exception at the University of Victoria (UVic) in Canada [2]. Typically, there is an attrition rate of more than 30% after the first two years in the Faculty of Engineering at UVic, which offers biomedical, civil, computer, electrical, mechanical, and software engineering programs. Attrition in our case includes students voluntarily dropping out of the program, being placed on probation, and failing out of the program.

At UVic, the first and second year of engineering programs provide the necessary fundamental skills and background to facilitate students' learning in higher level and more specialized courses. Therefore, the expected learning outcome of first year courses is for the students to master the necessary skills in mathematics and sciences in order to be successful in their programs. As well, some first-year students take this opportunity to find out more about the engineering profession and to decide whether it is a suitable career for them.

Many educators believe that there are performance indicators which can be used effectively to formulate correction actions to improve the attrition rate. It has been a traditional wisdom at engineering schools that among students who did poorly in mathematics and physics courses during their first and second year, have a higher attrition rate. However, this seemingly highly probable conjecture has never been proven nor refuted using a rigorous scientific approach; at least not that the authors are aware of.

To identify performance predictors and establish early warning indicators, data mining and statistical analysis techniques are used to analyze students' academic records. In this initial attempt, we focused on electrical and computer engineering students and first-year academic record only.

II. ACADEMIC PERFORMANCE PREDICTION IN THE LITERATURE

Engineering education is an actively research area. Most of the work has been focusing on curriculum revision and teaching improvement.

Recently, the use of technology to improve education and learning has gained much attention. Romero and Ventura present a survey of educational data mining for the period of 1995 to 2005 [9]. They find that most projects are targeted towards improving student learning activities, instructor teaching methodologies, and institution structuring. The same authors further introduce the use of data mining in predicting student performance in a course within the context of e-learning and intelligent tutoring systems [10].

Nghe, Janecek, and Haddaway use decision trees and Bayesian Network algorithms to predict a student's third-year GPA using the student's second year record [6]. However, they have not identified factors that effect success or failure; hence, their techniques cannot be of further use in improving student performance.

Azmi and Paris use similar techniques as Nghe et al. to predict and classify students into groups of various academic standing, based on student records [1]. This classification though lacks the identification of the relevant predictors of success, and simple lumps a student's complete degree record in its analysis.

978-0-7695-4992-7/13 $26.00 © 2013 IEEE
DOI 10.1109/CISIS.2013.15

Most of the publications in education performance prediction and data mining deal with e-learning and tutoring system using artificial intelligence techniques [7][11]. In existing literature, to the best of our knowledge, there is no specific work that examines and identifies performance predictors based on a student's academic record, similar to the work reported here.

III. Data Collection and Preparation

A student's record is available in a semi-structured text document. This text document is the same as the one a student can view and download through a web browser. Even though this is a proof-of-concept work, we tried to be flexible so that any text document can be utilized without the need to deal with specific database structure used by any university or organization. Figure 1 shows part of a sample transcript in free text format as downloaded.

The student record simply shows the courses taken and grades obtained, as well as the GPA and program standing (i.e., good, probationary, failed). There might be other factors that affect a student's performance but as a first attempt, the information as shown in Figure 1 is sufficient for some exploratory analyses.

To extract relevant information, a parser, written in Python [8], has been developed to recognize key words from the text-based record. A database, in SQLite [13], has been designed to accommodate the information retrieved. The database interface is implemented as a separate module to facilitate migration to a different database management system. Figure 2 shows the schema of this database. The attributes of the schema are used individually and in combination for later mining process.

A. Confidentiality and Privacy

It is important that one can view the database and an individual's record without the ability of identifying the individual. To this end, a secondary database is constructed during the information retrieval process.

Student names and identification numbers are extracted by the parser but they are kept in a separate secure database. As shown in Figure 3, an auto-generated 'public id' is used to map the student number to the corresponding record in the primary database. This way, one can explore the anonymous data without violating any confidentiality or privacy issues. Though, designated personnel can trace a public id back to the individual in the secondary database, in cases where advising activities are desirable for that individual.

B. Attributes Considered

For feasibility illustration, all first year courses are selected as attributes in the feature set for exploration. The feature set, or any information retrieved from the primary database, is queried into a CSV (Comma-Separated Values) format [12] for its compatibility with a large number of software packages such as Excel [5] and MATLAB [4].

Over the past 5 years, there were some changes to the first-year curriculum at UVic (all students in different disciplines share a common first year). For examples, merging of two related courses into one single course, splitting a course into two easily manageable subjects, and renaming of courses to reflect content changes. All these changes are consolidated and reflected in the feature set as shown in Table 1 which shows the features (i.e., course designation and number) and their titles. Equivalent courses are also listed and signified by a slash (/).

C. Dataset For Exploration

The dataset used for exploring performance predictors contains student records as represented by the feature set. For each student, it is necessary to quantify the 'performance' parameter to make the mining results more meaningful.

The students are classified into one of the three categories:

'S': Successful for students who obtained their engineering degree without ever placed on probation.

'P': Probation for students who have graduated but at one point in time were placed on probation indicating that they performed marginally.

'F': Failed for students who were required to withdraw from the engineering program and therefore no engineering degree was granted.

After reviewing the dataset, it is found that there was another category 'I' – in progress, for those students who have not failed but also have not received their engineering degree. These students, who eventually will fall into one of the above three categories, were removed from the dataset as it is meaningless to include them since their future outcome is not known yet.

The parser is capable of extracting the standing status of a student from the record and tagging a label of S, P, F, or I to the student. The process of determining the status category is shown in Figure 3. We believe this classification process gives a clear overall picture of a student's status and possibly his/her future.

Of the 90 student records collected, 29, 21, 30, and 10 are in the S, P, F, and I category, respectively. The 10 'I' records are thrown away with 80 remaining. It is then noticed that only a total of 72 records are usable since the other 8 are transferred students from other colleges, and therefore they do not have the first-year courses feature set required.

The eventual dataset used for exploration has 26 S, 16 P, and 30 F students.

IV. Principal Components

In our approach, each feature set or vector has thirteen features. It makes sense to first identify the important features that have larger impacts on the dataset. A commonly used tool is Principal Component Analysis (PCA) [3]. PCA is a dimensionality reduction technique and

provides weighted features in the form of main principal components. One drawback of PCA is that it only retains about 96% of the original information and ignores the outliers. However, this is not a concern in our case as we are looking for common trends rather than out of ordinary cases.

Using PCA on our dataset, the top ten principal components contribute to the total variance are shown in Table 2. The first principal component covers the majority of the information in the dataset at 86%. The features, in our case courses, with the largest influence on the first principal component would have the largest impact on the student record, and hence, the student's performance or classification.

It is worthwhile to examine each feature/course's contribution to the first principal component. As shown in Table 3, it can be concluded that the top three courses that have the most impact on a student's performance are MATH 100, MATH 110/MATH 133, and ENGR 120. A manual check on a sample of students using these top three indicators to predict performance concurred with our PCA findings.

V. DISCUSSIONS

Mathematical skills are highly relevant to an engineer's work. Those lacking or deficient in such skills would ultimately have difficulty in completing their engineering degree. It makes sense that the top two indicators of a student performance are mathematics courses.

Somewhat unexpected, or rather this has never been considered, is the third course that influences highly a student's performance. ENGR 120 has two components: technical English and introduction to engineering design. Design and Communication II is considered by students to be an easy and soft course. Though, a student's grade in this course is indicative of future academic performance.

To promote retention, our Faculty has several initiatives to assist students in mastering their mathematical subjects. It has also been recommended that the course content and delivery need to be reviewed and possibly revised for ENGR 120.

VI. FUTURE WORK

Using PCA to identify academic performance has shown potential and promise. It is worthwhile to further pursue the use of data mining and statistical analysis techniques to pinpoint problematic courses that affect a student's performance in later years of their study. Corrective actions can then be carried out to improve retention rate.

We are investigating both supervised and unsupervised learning techniques to explore academic performance prediction. These techniques include various clustering and classification approaches, such as K-means and hierarchical clustering, and K-nearest neighbor and naïve Bayes classifiers. In addition, each of these approaches' effectiveness and accuracy will be quantified.

In this work, the features considered are the grades obtained in the thirteen first-year courses. There might be other variables or of their combination, such as age, gender, and mother-tongue, that are highly indicative of a student's performance. We plan to carry out further experiments using more features to explore the academic performance space.

REFERENCES

[1] Azmi, M., and I. Paris, "Academic performance prediction based on voting technique", IEEE International Conference on Communication Software and Networks, pp. 24-27, 2011.

[2] Faculty of Engineering, University of Victoria, British Columbia, Canada. URL: http://www.engr.uvic.ca [accessed: April 1, 2013].

[3] Jackson, J.E., *A User's Guide to Principal Components,* John Wiley & Sons, Inc., 2003.

[4] MATLAB, MathWorks. URL: www.mathworks.com [accessed: April 1, 2013].

[5] Microsoft, Office Excel. URL: http://office.microsoft.com/en-ca/excel/ [accessed: April 1, 2013].

[6] Nghe, N.T., P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance", 37th ASEE/IEEE Frontiers in Education Conference, pp. T2G 7-12, 2007.

[7] Pardos, Z. et al. "The effect of model granularity on student performance prediction using Bayesian Networks", LNAI 4511, pp. 435-439, 2007.

[8] Python, Computer Programming Language. URL: http://www.python.org [accessed April 1, 2013].

[9] Romero, C, and S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications 33, pp. 135-146, 2007.

[10] Romero, C, and S. Ventura, "Educational data mining: A review of the state-of-the-art", Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, Pre-print, 2012.

[11] Romero, C., S. Ventura, and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers and Education, Pre-print, 2012.

[12] Shafranovich, Y., "Common Format and MIME type for Comma-Separated Values (CSV) Files", Internet RFC 3987, October 2005. URL: http://www.ietf.org/rfc/rfc4180.txt [accessed: April 1, 2013].

[13] SQLite, Open Source Database. URL: http://www.sqlite.org [accessed April 1, 2013].

```
WINTER 2005-2006
  ENGINEERING B.ENG.
  (CO-OP ENGINEERING)
      CHEM  150   ENGINEERING CHEMISTRY   1.50  B     4     1.50
      CSC   110   FUNDAMENTAL PROGRAMING:I      1.50  F     0     0.00
      CSC   110   FUNDAMENTAL PROGRAMING:I      1.50  B+    5     1.50
      ELEC  199   LAB: ENGR FUNDAMENTALS  1.00  B+    6     1.00
      ENGL  135   ACADEMIC READING+WRITING      1.50  B-    5     1.50
      MATH  100   CALCULUS:I  1.50  C     2     1.50
      MATH  101   CALCULUS:II       1.50  F     0     0.00
      MATH  133   MATRIX ALG FOR ENGINEERS      1.50  F     0     0.00
      MECH  141   ENGR FUNDAMENTALS:I     1.50  E     0     0.00
      PHYS  122   MECHANICS FOR ENGINEERS       1.50  C     3     1.50
      Credit in 8.50 Units
      Sessional GPA = 2.53
      PLACED ON FACULTY PROBATION
      MECH  141   SUPPLEMENTAL      1.50  B     1     1.50
      Credit in 1.50 Units
      PROGRAM STANDING: PROBATIONARY 29 APR 2006
```

Figure 1. A Sample Text Based Student Record

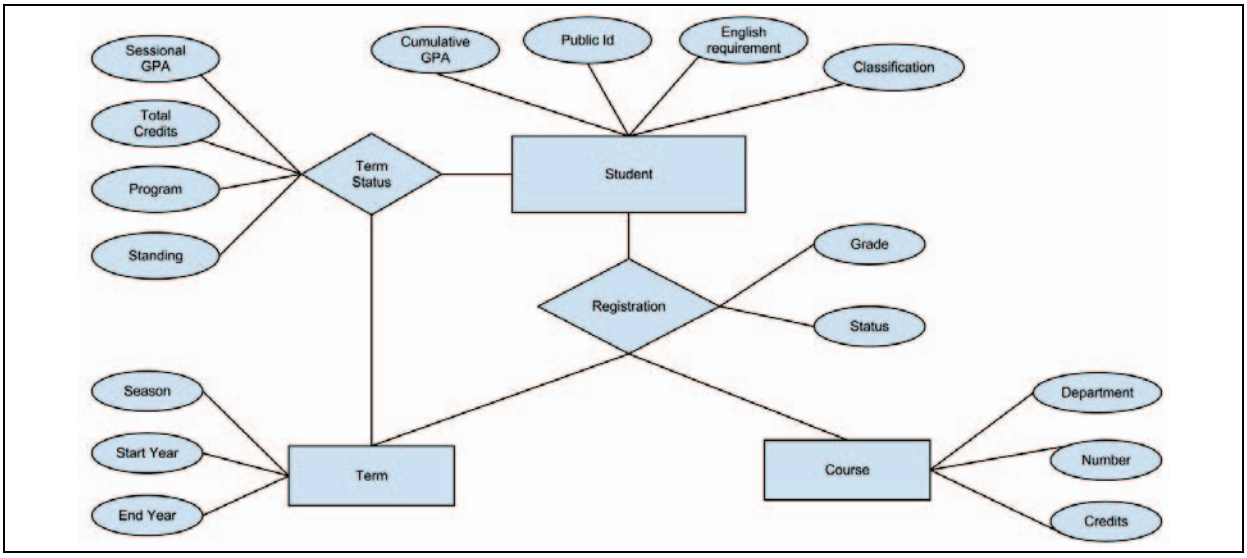

Figure 2. Database Schema Representing Student Record

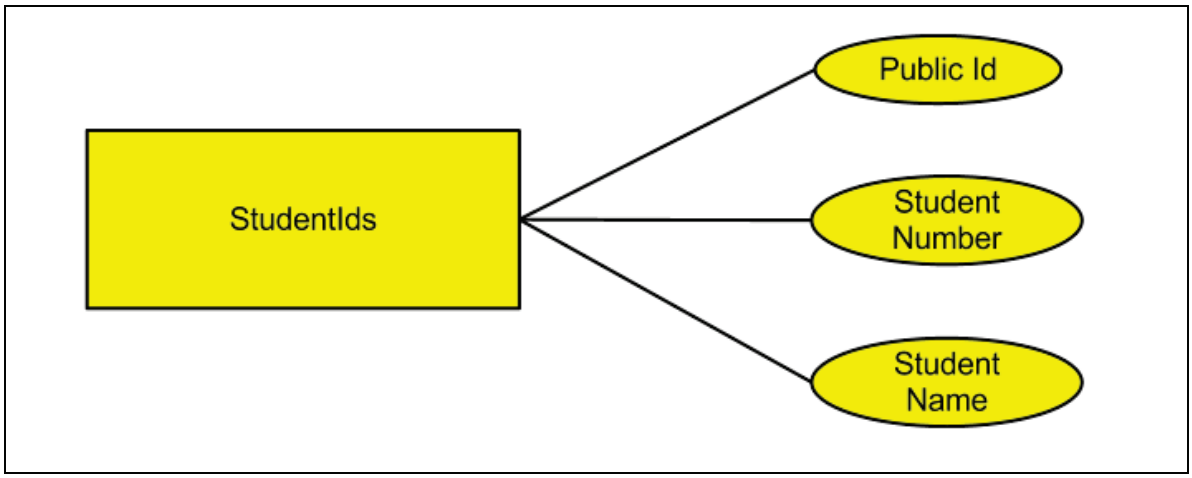

Figure 3. Security Mapping Between Primary and Secondary Database

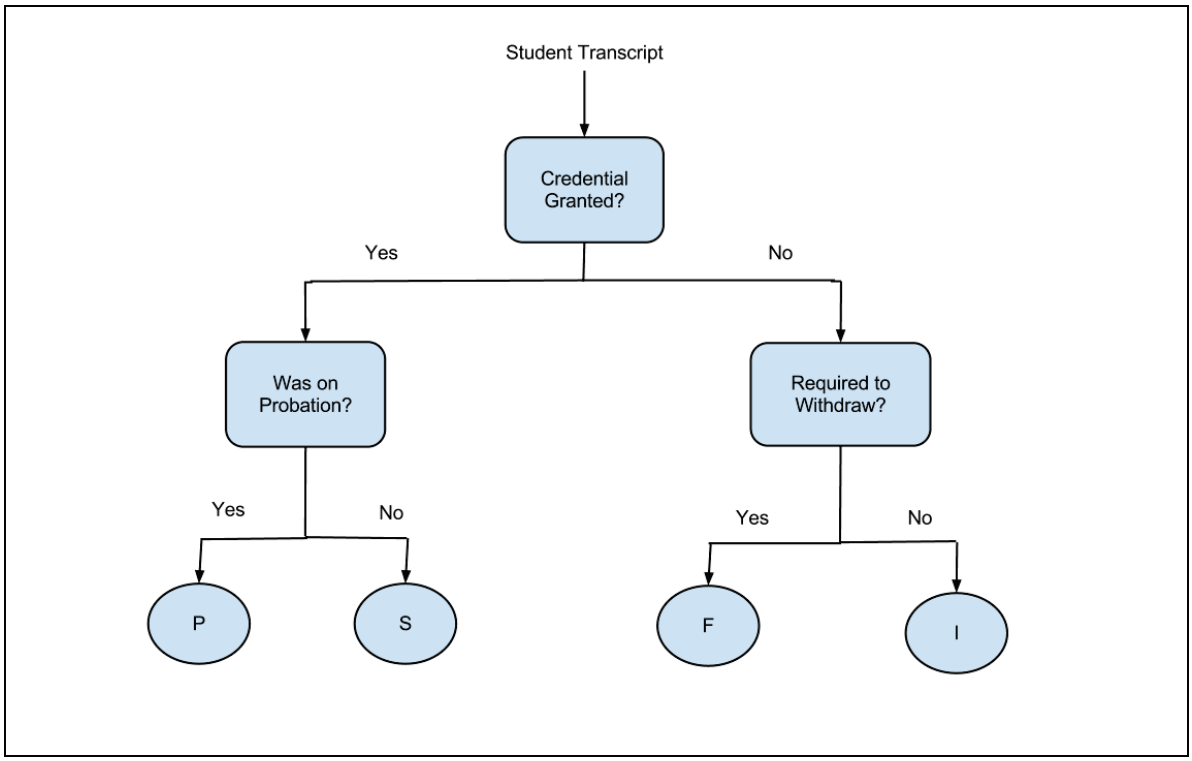

Figure 4. Determining a Student's Status

Table 1. Feature Set For Exploration

Feature	Title
ELEC 199	Laboratory in Engineering Fundamental
CHEM 150	Enginering Chemistry
PHYS 125	Fundamental of Physics
CSC 115 / CSC 160	Fundamental of Programming II
ENGR 110	Design and Communication I
CSC 111 / CSC 110	Fundamental of Programming with Engineering Applications
MATH 100	Calculus I
MATH 101	Calculus II
ENGL 135 / ENGL 115	Academic Reading and Writing
MATH 110 / MATH 133	Matrix Algebra for Engineers
ENGR 120	Design and Communication II
PHYS 122	Mechanics for Engineers
MECH 141 / ENGR 141	Engineering Fundamentals I

Table 2. Contribution to the Total Variance by Each Principal Component

Principal Component	Percentage of Total Variance
1	86%
2	3%
3	2%
4	2%
5	2%
6	1%
7	1%
8	1%
9	1%
10	1%

Table 3. Each Course's Contribution to the First Principal Component

Course	Magnitude of Contribution
MATH 100	0.19138
MATH 110 / MATH 133	0.18792
ENGR 120	0.17206
ELEC 199	0.13868
MATH 101	0.12529
CHEM 150	0.12335
PHYS 125	0.12010
CSC 115 / CSC 160	0.11692
CSC 111 / CSC 110	0.09954
MECH 141 / ENGR 141	0.07175
ENGL 135 / ENGL 115	0.03726
PHYS 122	0.03009
ENGR 110	0.02447