Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program

Pimpa Cheewaprakobkit

***ABSTRACT*--The aim of this study is to analyze factors affecting academic achievement that contribute to the prediction of students' academic performance. It is useful in identifying weak students who are likely to perform poorly in their studies. In this study, the researcher used WEKA open source data mining tool to analyze attributes for predicting undergraduate students' academic performance in an international program. The data set comprised of 1,600 student records with 22 attributes of students registered between year 2001 and 2011 in a university in Thailand. Preprocessing included attribute importance analysis. The researcher applied the data set to differentiate classifiers (Decision Tree, Neural Network). A cross-validation with 10 folds was used to evaluate the prediction accuracy. An experimental comparison of the performance of the classifiers has been conducted. Results show that the decision tree classifier achieves high accuracy of 85.188%, which is higher than that of neural network classifier by 1.313%.**

***Index Terms*--Academic Achievement, Data Mining, Decision Tree, Neural Network**

I. INTRODUCTION

Higher education is an important contributor to the development of human resources. However, one of the major problems is failure in graduate education. There are students with a GPA below the required standard. As a result, students cannot graduate in a given period of time and lose job opportunities. Each year the number of students who are dropping out has been growing. Therefore, this research is aimed to investigate the factors that affect the academic achievement of students. The data sample used in this study was a group of undergraduate students in an international program and a group of students at risk in their academic performance that is below a certain threshold as measured by cumulative grade point average (CGPA) of less than 2.00. The ultimate aim is to advise and assist the students in a timely fashion. This analysis uses data mining techniques to classify the data.

Classification is the process of data management model building that identifies in-group data to illustrate the differences between groups of data and to predict the data that should be in any class. The model used to classify data into determined groups is based on an analysis of the data set. This data set would lead the system to classify data.

Manuscript received September 7, 2012; revised October 15, 2012. This work was supported in part by Asia-Pacific International University.

Pimpa Cheewaprakobkit is with the Computer Information System Department, Faculty of Business Administration, Asia-Pacific International University, P.O. Box 4, Muak Lek, Saraburi 18180, Thailand. (Phone: 66-36-720777; e-mail:pimpa@apiu.edu)

The end result is a model of learning which can be represented in many forms such as the Classification (IF-THEN) rules, Decision Tree, or Neural Networks. Then the rest of the data, as the actual data, will be drawn to test and compare with those acquired from the model for the accuracy testing. The model will be updated and tested to have a satisfactory level. Later, when new data comes and is plugged into the model, the data can predict grouping by the model.

II. THEORIES AND RELATED WORKS

A. Data mining

Data mining [7] is a process of automatically discovering useful information in large data repositories. It is an integral part of Knowledge Discovery in Database (KDD), which is the overall process of converting a series of transformation steps, from data preprocessing to the post-processing of data mining result. Data mining tasks are generally divided into 2 major categories, namely, predictive and descriptive tasks. Predictive modeling refers to the task of building a model for target variable as a function of the explanatory variable. The two types of predictive modeling tasks are classification, which is used for predicting discrete attributes and regression, which is used for predicting continuous target attributes. The goal of both tasks is to create a model that minimizes the error between the predicted and true values of the target variable.

B. Related Works

Work by Thai Nghe, Janecek, and Haddawy [11] have compared two classifiers (Decision Tree and Bayesian Network) to predict students GPA at the end of the third year of undergraduate studies and at the end of the first year of postgraduate from two different institutes. Each data set has 20,492 and 936 complete student records respectively. The results show that the Decision Tree outperformed Bayesian Network in all classes. The accuracy was further improved by using resampling technique, especially for Decision Tree, in all cases of classes. At the same time, resampling was used to reduce misclassification, especially on minority class of imbalanced datasets, because Decision Tree algorithm tends to focus on local optimum. Ian and Eibe [4] gave a case study that used educational data mining to identify behavior of failing students to warn students at risk before final exam. Romero, Ventura and Garcia [12] gave another case study of using educational data mining in Moodle course management system. They used each step in data mining process for mining e-learning data. Also,

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

educational data mining used by Polpinij [5] to predict students' final grade using data collected from Web based system. Beikzadeh and Delavari [1] used educational data mining to identify and then enhance educational process in higher educational system which can improve their decision making process. Finally, Waiyamai [14] used data mining to assist in development of new curricula, and to help engineering students to select an appropriate major.

Other works, Kotsiantis, Pierrakeas, and Pintelas [13] have compared six classification methods (Naive Bayes, Decision Tree, Feed-forward Neural Network, Support Vector Machine, 3-nearest Neighbor and Logistic Regression) to predict drop-outs in the middle of a course. The data set contained demographic data, results of the first writing assignments and participation in group meetings. The data set contained records of 350 students. Their best classifiers, Naive Bayes and Neural Network, were able to predict about 80% of drop-outs. The results also showed that a simple model such as Naïve Bayes is able to generalize well on small data sets compared to other methods such as Decision Tree and Nearest Neighbor that require a much larger size of datasets.

III. MATERIALS AND METHODS

To investigate the propositions, two classification algorithms have been adopted and compared: the neural network and the C4.5 decision tree algorithm. The classification models were implemented using WEKA 3.7.5 version [9]. Series of records of first year undergraduate students who enrolled in the international programs of a private university in the academic years of 2001 -2011 with 1,600 items and 22 attributes were used for the study. The investigation process consists of three main steps [10]: Data Preprocessing; Attribute Filtering; and Classification Rules.

A. Data Preprocessing

The student records were still not in a form that could be used in the data mining testing and analysis: therefore, the data needed to be prepared to be in the proper format before using them. The process was divided into various stages: Data Cleaning; Data Selection; and Data Transformation. The records of samples were drawn from many departments: for example, the study performance samples were taken from the Office of Admissions and Records, the number of hours in the extra curricula activities was taken from the Office of the Student Administration, and the number of hours worked was taken from the Finance Department. These data were in the form of several Microsoft Excel files with some duplicate fields. To make it easier to write programs, the researcher restored the data into a form of table using the Oracle Database version 10g Express Edition as shown in Fig 1. The program was developed by Java and SQL language for selecting the attributes as presented in Fig 2. Then, the researcher recalculated values of the attribute. For example, the researcher recalculated the number of hours worked and the number of hours in the extra curricula activities per semester to per month, and from the hours of study outside classroom per semester to each month for all the studied students.

Then, the researcher collected all data and Export File from the table records format into the .CSV data file format as shown in Fig 3, to be used in the analysis with the WEKA. The attributes used in research were reduced into only the desired attributes.

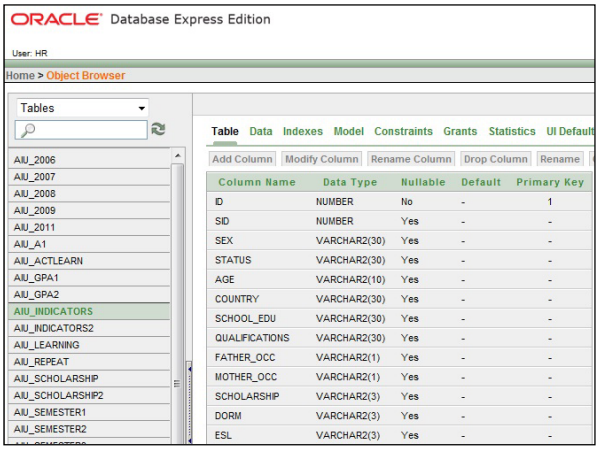

Fig 1. The data import from Excel file into the Oracle database

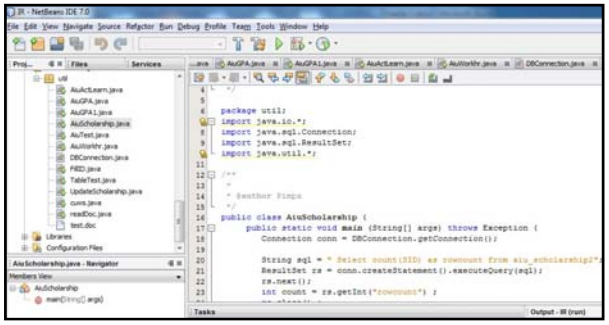

Fig 2. The development of program through Java language to select desired attributes

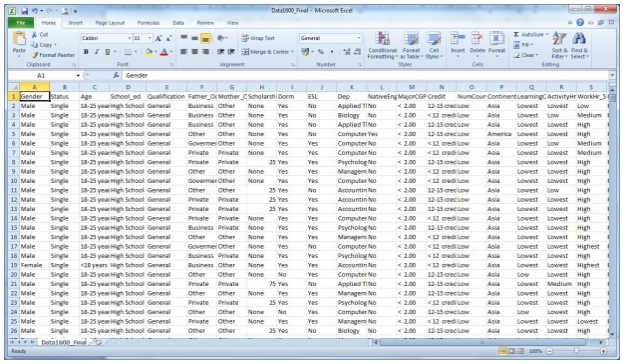

Fig 3. Example of students' data after the cleaning

B. Attribute Filtering

This is the process of screening for the most needed attributes. The working principle of data classification of this study was based on the selected 20 attributes out of the initial 22 attributes as shown in Table 1. The student Identification Number (ID) and semester Grade Point

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Average (GPA) were left out from the study through this process. The student Identification Number for all students, were either identical or different from each other and did not affect the analysis. Because the semester GPA of the students, was close to Cumulative Grade Point Average (CGPA), the CGPA was selected as the dependent variable in this analysis. If the GPA attribute was included in this study, it would have the most influence in the study and the relationship of other attributes would not be seen.

C. Classification Rules

Classification of data is a technique for data classification from various feature items through the survey of attributes in the database to distinguish categories which have been defined in advance. The techniques used to categorize the data include the Decision Trees and Neural Network techniques

A Decision Tree is a tree-shaped structure that represents sets of decisions [2]. These decisions generate rules for the classification of a dataset. Trees develop arbitrary accuracy and use validation data sets to avoid spurious detail. They are easy to understand and to modify. Moreover, the tree representative is more explicit, and has easy-to-understand rules for each cluster of student's performance. The classes in the Decision Tree are cluster IDs obtained in the first step of the method. The Decision Tree represents the knowledge in the form of IF-THEN rules. Each rule can be created for each path from the root to a leaf. The leaf node holds the class prediction.

The C4.5 is an algorithm used to generate a Decision Tree developed by Ross Quinlan. The C4.5 is an extension of Quinlan's earlier ID3 algorithm. It employs a "divide and conquer" strategy and uses the concept of information entropy. The general algorithm for building Decision Trees is: [6]

- If all cases are of the same class, the tree is a leaf and so the leaf is returned labeled with this class;
- For each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Also calculate the gain in information that would result from a test on the attribute (based on the probabilities of each case with a particular value for the attribute of a particular class)
- Depending on the current selection criterion, find the best attribute to branch on.

A multilayer perceptron [8] is a Feed forward Artificial Neural Network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable by a hyper plane. MLP networks are general-purpose, flexible, nonlinear models consisting of a number of units organized into multiple layers. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has been shown that MLPs can approximate virtually any function to any desired accuracy.

Table I: All variables in the student database

Attribute	Type	Description
ID		student's identification
Gender	binary	student's gender (female or male)
Status	nominal	student's status (single, married, divorced)
Age	numeric	student's age (1:< 18 years, 2: 18-25 years, 3: 26-30 years, 4: 31-40 years, 5: > 40 years)
Continent	nominal	Continent (Asia, America, Europe, Africa)
School_Ed	nominal	Educational background (BA, College, Diploma, High School, MA)
Qualification	numeric	Student's qualification (1-General, 2-Vocational)
Father_Occ	numeric	Father's occupation (1-Government, 2-Private, 3-Business, 4-other)
Mother_Occ	numeric	Mother's occupation (1- Government, 2- Private, 3-Business, 4- Other)
Scholarship	numeric	Get a scholarship (25%, 50%, 75%, 100%)
Dorm	binary	On-campus residence (yes or no)
ESL	binary	Pre-University English ESL (yes or no)
Dept	numeric	Department (from 1 – 10)
NativeEng	binary	Native English speaker (yes or no)
G.P.A	numeric	Grade Point Average (< 2.00, 2.00-2.49, 2.50- 2.99, 3.00-3.49, > 3.50)
C.G.P.A	numeric	Cumulative Grade Point Average (< 2.00, 2.00-2.49, 2.50-2.99, 3.00-3.49, > 3.50)
Major_CGPA	numeric	Grade Point Average of major subject(< 2.00, 2.00-2.49, 2.50-2.99, 3.00-3.49, > 3.50)
Credits	numeric	Number of credits (from 1 – lowest to 3 – highest)
NumCourse	numeric	Extra-curricular subject (from 1 – lowest to 3 – highest)

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Attribute	Type	Description
Learning_Center	numeric	Extra study hour (from 1 – lowest to 5 – highest)
Work_Hour	numeric	Work hour (from 1 – lowest to 5 – highest)
Activity_Hour	numeric	Work activity (from 1 – lowest to 5 – highest)

D. Model Building

For the data classification format determinant for the Training Set and the set of data format to use in testing the validity of the Testing Set, the researcher used the classification technique of Cross-validation Fold: 10 [11]. This method divided the data into 10 sets and at each time of study, one data set was used for testing and the remaining nine sets were used to develop the model. The testing was repeated to cover the 10 series data set. Then, the study was tested several times by adjusting the values, choosing the Correct Value, and comparing the Precision Value and the Recall Value to have the most appropriate value to be used for the model.

E. Model Evaluation

To create a rule of Decision Trees to use as a model, the researcher selected the technique of Decision Rules: PART by selecting rules with a clear condition: not too many or too few rules: rules that can be easily understood: and the appropriate Correct Value, Precision Value and Recall Value. However, whenever there were too many rules, the researcher used the pruning method to reduce the classification errors caused by outliers, and then compared the tested results with the Neural Network method from the Correct Value, Precision Value, and Recall Value.

IV. RESULTS

The results of the analysis show the Decision Tree Model had an accuracy rate of 85.188% while the Neural Network Model had an accuracy rate of 83.875%. The result suggests that the Decision Tree Model is more accurate than the Neural Network Model. Further results reveal the factors that affect academic achievement of students are as follows:

1. The number of hours worked per semester;
2. An additional English course;
3. The number of credits enrolled per semester;
4. Status of students such as single, married, or divorced.

The results from the test created 13 rules. Following are examples of these rules:

WorkHr_5 = High: Risk (320.0/12.0)

WorkHr_5 = Low AND
Credit = 12-15 credits AND
Mother_Occ = Other: Not Risk (65.0/7.0)

ActivityHr_5 = Medium: Risk (9.0/2.0)

ESL = No AND
Credit = 12-15 credits: Not Risk (290.0/50.0)

NumCourse = Low AND
Continent = Asia AND
Age = 18-25 years AND
ActivityHr_5 = Low AND
WorkHr_5 = Medium: Risk (14.0/3.0)

Credit = 12-15 credits AND
ESL = No AND
ActivityHr_5 = Lowest AND
Status = Single AND
School_ed = High School AND
Dorm = Yes: Not Risk (82.0/21.0)

WorkHr_5 = Lowest AND
NumCourse = Low: Not Risk (139.0/39.0)

Credit = < 12 credits AND
Gender = Male AND
Status = Single: Risk (95.0/18.0)

ESL = No AND
Credit = > 15 credits: Not Risk (109.0/9.0)

Table II: Performance comparison between Decision Tree and Neural Network models

Performance Measures	J48 (C4.5)	MLP
Correctly Classified Instances (%)	85.188	83.875
Incorrectly Classified Instances (%)	14.812	16.125
Precision	0.852	0.838
Recall	0.852	0.838

Table II shows performance comparison between the Decision Tree and the Neural Network models. When comparing the Precision Value and the Recall Value of the 2 models, the Decision Tree model generates the Precision Value of 0.852. This represents the number of found class and a prediction accuracy of 85.2% when compared to the number of whole classes from the database. It is in line with the Recall Value of 0.852 which also means the number of found class and a prediction accuracy of 85.2% when compared to the number of whole classes from the database. On the other hand, the Neural Network model generates the Precision Value of 0.838 or 83.8% and the Recall Value of 0.838 or 83.8%. The results reveal that the Decision Tree Model gives more accurate prediction than the Neural Network Model.

When comparing the Precision and Recall values of the Decision Tree Model, the results are equal at 0.852 or 85.2%, while the Precision and Recall values of the Neural Network Model are also equal at 0.838 or 83.8%. The results show the accuracy of prediction when compared to the found classes and all classes in the database. The percentages of accuracy are equal.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. CONCLUSION AND FUTURE WORK

After analyzing the factors affecting student achievement of undergraduate students registered in the International Program using the Decision Tree and Neural Network techniques, it can be concluded that the Decision Tree technique has better efficiency data classification for this data set.

The analysis of important factors for grouping students could be concluded as follows:

Firstly, most of the students who do not have risk of low academic achievement are the students who have never studied additional English courses. The result shows that this group of students had a good base of English proficiency before entering the university. This is why they did not need to take additional English courses to improve their English skill. They are single, work few hours per semester, and register for 12-15 credits per semester.

Secondly, most of the students who have risk of low academic achievement are the students who study additional English courses. The result shows that this group of students did not have good base of English proficiency before entering the university. This is why they needed to take additional English courses to improve their English skill. Many of them are either married or divorced. They work at a moderate to high number of hours per semester. They register for either less than 12 credits per semester (students are not allowed to register more than 12 credits if the CGPA is low) or more than 15 credits per semester (students are allowed to register more than 15 credits if the CGPA is high).

There were several limitations to the study and any future study should expand to look at some of the following issues. Firstly, the results from the data analysis from the data mining method are only the important factors affecting student achievement. Each factor has a different significant value. Thus, the grouping of students who are at risk or not at risk, and other factors or elements should be considered as well. Secondly, the model can be improved to be applicable to analyze the risk level of students, and find ways to advise and assist the at-risk students. Thirdly, the research analyzed only data from students registered in the international program of the undergraduate level. Future investigation should expand the study to analyze the students' performance in other programs.

There are some recommendations to the institution studied. Firstly, there should be a system to record important student information accurately and completely. Lastly, there should be a central database to store the information of all students.

REFERENCES

[1] M. R. Beikzadeh and N. Delavari, "A New Analysis Model for Data Mining Processes in Higher Educational Systems," On the proceedings of the 6th Information Technology Based Higher Education and Training 7-9 July 2005.

[2] T. Connolly, C. Begg and A. Strachan, "Database Systems: A Practical Approach to Design Implementation and Management," Harlow: Addison-Wesley, 1999, pp.687.

[3] G. Piatetsky-Shapiro and W. J. Frawley, "Knowledge Discovery in Databases," MIT Press, 1991.

[4] H. W. Ian and F. Eibe, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," California: Morgan Kaufmann, 2005.

[5] J. Polpinij, "The Probabilistic Models Approach for Analysis the Factors Affecting of Car Insurance Risk," M.S. thesis, Department of Computer Science, Kasetsart University, Thailand. 2002.

[6] J. Quinlan, "C 4. 5: Programs for Machine Learning," Morgan Kaufmann, 1992.

[7] L. S. Affendey, I.H.M. Paris , N. Mustapha, Md. Nasir Sulaiman and Z. Muda, "Ranking of Influencing Factors in Predicting Students' Academic Performance," International Technology Journal, vol. 9, no. 6 , pp. 832-837, 2010. ISBN 1812-5638.

[8] M. T. Hagan and M. B. Menhaj, "Training Feed-forward Networks with the Marquardt Algorithm," IEEE Trans. on Neural Networks, vol. 5, no. 6, pp. 989-993, Nov. 1994.

[9] A. Merceron and K. Yacef , "Educational Data Mining: a Case Study," In Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press. 2005.

[10] B. Minaei-Bidgoli, D. Kashy, G. Kortemeyer and W. Punch, "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System," In the Processing of 33rd ASEE/IEEE conference of Frontiers in Education. 2003.

[11] N. Thai Nghe, P. Janecek, and P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance," ASEE/IEEE Frontiers in Education Conference, 2007.

[12] C. Romero, S. Ventura and E. Garcia, "Data Mining in Course Management Systems: Moodle Case Study and Tutorial," Computers & Education, vol. 51, no. 1. pp. 368-384. 2008.

[13] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing Student Dropout in Distance Learning Using Machine Learning Techniques," In proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003), pp. 267-274, 2003. ISBN 3-540-40803-7.

[14] K. Waiyamai, "Improving Quality of Graduate Students by Data Mining," Department of Computer Engineering, Faculty of Engineering, Kasetsart University , Thailand. 2003.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)