The Golden Teak : Humanity and Social Science Journal (GTHJ.)

Vol.22 Special Issue September - December 2016 ISSN 2408 - 0845

# A Study of Factors Influency Student Dropout Rate Using Data Mining

Butsaraporn Mahatthanachai*

Hathaithip Ninsonti**

Nuttiya Tantranont***

## ABSTRACT

The objective of this research is to study factors affecting student dropout rate using data mining. The information of students in the Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University was used as data in this study. The data were divided into two sets. The first set was obtained from former institution containing 1,433 records of data. This data set was analyzed with classification data-mining technique based on decision tree and C4.5 algorithm. The analysis results showed that the top three factors that have impact on dropout of the students are previous study result (GPA), previous study field, and educational background, with accuracy values of 72.02, 70.11, and 68.13, respectively. The second data set contains 2,568 records of current study results. This set was used to determine courses that affect dropout of the students based on association rules data mining technique and Apriori algorithm. The courses found to have impact on the students' dropout were Computer, English, Mathematics and Physics. Results from this research can be useful for developing a prediction model of student dropout. In addition, the research findings can provide some guidelines for universities to solve the dropout problem of their students.

**Keyword : Student Dropout / Data Mining / Decision Tree / Association Rules**

*Ph.D Candidate, Asian Development College for Community Economy and Technology, Chiang Mai Rajabhat University

**Lecturer, Asian Development College for Community Economy and Technology, Chiang Mai Rajabhat University

***Corresponding Author, Lecturer, Asian Development College for Community Economy and Technology, Chiang Mai Rajabhat University

## Inroduction

Higher educational institutions are responsible for producing graduates who have quality, skills, and potential to apply their gained knowledge usefully with their appropriate occupations according to the national economic and social development plan. Major components of the process to enhance quality of students include wisdom, family background, learning trait, etc. These factors are involved directly with quality and achievement of students. The 11th Higher Education Development Plan (2012-2016) has set a goal to have students graduate within 4 years for 70%. However, some of students in the higher educational level have to retire or be retired during their study. This dropout causes some damage to the institution in terms of administration and investment in resources. Moreover,the retired students also lose time and money. Therefore, this is a major problem that challenges many higher educational institutions around the world. In Thailand, many universities encounter the problem of student dropout. For instance, Huachiew Chalermprakiet University had a ratio of retired students for 25.23% of all students (Thongkon, 2013). Similarly, King Mongkut's University of Technology North Bangkok had students with dropout for 26.29% of the total students (President’s Offices, 2011). Research results regarding dropout of students in various institutions reveal that there are multiple causes including learning efficiency of the student (Hsieh, Sullivan & Guerra, 2007), expectation of the student (Chemers, Hu, & Garcia, 2001; Lee, Kang, & Yum, 2005), learning outcomes from the previous institution (Bean, 1980), financial problem (Mohr, Eiche, Sedlacek, 1998; Bean,1980), problems of the institution such as a lack of educational strategy (Chemers, Hu, & Garcia, 2001; Lee, Kang, & Yum, 2005), and student's adjustment to the university (Willcoxsona, Cotterb & Joyc, 2011). In addition, some students were found to have insufficient knowledge and understanding about the curriculum, courses, courses selection, registration method, learning approach, learning assessment, requirements, and regulations of the university.

Chiang Mai Rajabhat University is another institution that faces the dropout problem. Generally, the University sets a targeted number of graduates to be produced in each year based on its budget. However, in practice, it could not produce graduates according to the goal. This situation affects the budget that it is supposed to be supported by the government, and also affects its quality in yearly graduate production. Data from the Office of the Registrar, Chiang Mai Rajabhat University showed that the dropout rate of students studying during 2007-2015 was 14.8%. In each academic year, many students have retired. Thus it is important for the University to find causes of the dropout so that the finding can provide a guideline for improving quality of teaching and learning.

This research is aimed to find factors affecting student dropout rate via the use of data mining technique. The study was based on data of students in the Department of Computer, Chiang Mai Rajabhat University. The research result could be used to create a model for prediction of students dropout. This model will also be useful in warning students to adjust their learning methods in the remaining courses for improving their grade point average (GPA) so that they can pass the criterion at the end of each semester.

**Objectives**

The main objective is to study factors influencing students dropout by using a data mining technique as a case study of students in the Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University.

**Methodology**

1. Study of general data of Chiang Mai Rajabhat University's students and some specific data of students in Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University, in order to examine some relevancy with dropout of the students that may found in the analysis process.

2. Analyze the collected data with the use of data mining technique

The general data and educational data of students were taken to study factors that have influence toward dropout of the students with the use of data mining technique. The analysis can be divided into two steps, as shown in Figure 1.

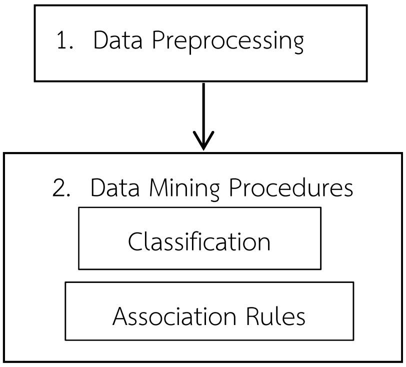

**Figure 1** Data analysis using data mining technique

2.1 Data Pre-processing

In this research, the data were collected from those students who have GPA records during the 2007-2015 academic years. There are 1,433 students in total. The data were divided into two parts as follows.

1) Data from previous academic institutions include these following attributes:

-Sex, which is classified into male and female

-Province of residence according to the province names in Thailand

-Educational level, which is classified into vocational level, higher vocational level, high-school education (upper Matthayomsuksa), and bachelor's degree.

-Fields of study, such as Science-Math field, Art-Language field, electronics, etc. There are 15 fields altogether.

-Grade point average (GPA), which is classified into 5 levels namely Excellent, Very good, Good, Moderate, and Insufficient.

-Occupation of parents, which is classified into 7 groups namely Farmers, Sellers, Civil servants, Government employees, State enterprise employees, General workers, and Miscellaneous.

-Student status, which is divided into 2 groups namely Current students and Retired students

2) Data about current study results were collected from students who had retired and got some E, F, D, and D+ grades. There are 2,568 records in total. The data contain these following attributes:

-Registered courses

-Study results

Data from previous institutions and current study results were screened with multiple procedures including data cleaning, data selection, and data transformation. Therefore, the data was appropriate for the analysis (Mahapatra & Bose, 2001). Details of the attributes are shown in Table 1 and some example data are shown in Tables 1.

**Table 1** Details of relevant attributes

| No. | Attribute Name | Type | Description |
|---|---|---|---|
| 1 | Sex | Text | Sex |
| 2 | Province | Text | Province of residence |
| 3 | Old_ED | Text | Previous education |
| 4 | Old_Major | Text | Previous academic field |
| 5 | Old_GPA | Text | Previous GPA |
| 6 | GD_Job | Text | Occupation of parents |
| 7 | Status | Text | Student status |
| 8 | Subject | Text | Registered course |
| 9 | Grade | Text | Current study result |

2.2 Data mining technique

The data mining technique was conducted in order to analyze the factors affecting students dropout. This analysis was performed by using software called WEKA (Waikato Environment for Knowledge Analysis). The analysis consists of two approaches namely classification and association Rules.

1) Classification

Classification is a procedure to take the Part 1 data (data of previous academic institution) to analyze the factors affecting students dropout. This procedure was done by using Decision Tree C4.5 algorithm (Pansumret, Phuboon-ob, & Pongsiri, 2013; Werghi and Kamoun, 2010.). Each branch represents attributes of the students, while each leaf node represents their status.

2) Association Rules

Association rules determination is done by taking the Part 2 data (current study results data) to determinine the factors affecting students dropout. Association rules were used to find association rules of grades for each course. The course and the corresponding grade were considered to be the same item from the value of Support ($\sigma$) and value of confidence ($\rho$) (Suwannarattaphoom & Waiyamai, 2012; Laokietkul & Sitthiworachart, 2008).

**Research Results**

1. The results of this study could be summarized with interesting topics as follows:

1.1 General data of students in Chiang Mai Rajabhat University

There are 37,174 students of Chiang Mai Rajabhat University who took Normal Section education during the academic years of 2007-2014. Among these students, the ratios of students who remain to have the current status as a student, graduated students, and retired students are accounted for 43.44%, 32.61%, and 23.94%, respectively (Office of the registrar, 2016). Based on the data, it was found that there are multiple causes that lead to failure of students to complete their education in each academic year. These causes include having no money for tuition fees, resigning, changing a major, having poor study results regarding criteria, moving to another institution, death, limitation of study period, and missing educational evidence from their previous academic institution.

Based on the studied data as discussed above, the two main causes related to study results that lead to the loss of student status are having poor study result under the criteria and limitation of study period. The retired students in each class year were studied and summarized as shown in Table 2.

**Table 2** Percentages of dropped out students during 2007-2010 academic years

| College year | Percentages of Dropout |
|---|---|
| 1 | 26.51 |
| 2 | 37.58 |
| 3 | 23.23 |
| 4 | 10.30 |
| More than 4 | 0.39 |

According to Table 2, the 2nd and 1st college years were found to have the highest ratios of dropped out students, respectively.

1.2 Specific data of students in the Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University

The data of students in the Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University during 2007-2015 academic years showed that there were 1,433 students who had registered for their study and got their study result records. The data could be categorized by programs as shown in Table 3.

**Table 3** Number of students in the Computer Department as categorized by programs

| Educational program | Number of students |
|---|---|
| Computer Science | 565 |
| Information Technology | 542 |
| Web Programming and Security | 326 |
| **Total** | **1,433** |

According to the data of students at department of computer, the total number of students are 1,433. The number of dropped out students as a result of academic grading falling below the standard are 202 (or 14.09% of 1,433 students). The data were then analyzed with the data mining technique to determine their association rules.

2. Data mining analysis

2.1 Classification analysis

The analysis to determine factors of interest with data classification approach was conducted by using the decision tree technique together with C4.5 algorithm to investigate the accuracy. Each branch represents attribute of the students, and each leaf node represents their status. The students could be categorized by their status into 3 groups namely current student status, dropped out student status, and other statuses such as having resigned, having not paid the tuition fee, etc. The data stored in the database contained 2,520 records, but only 1,433 records can truly be used. Model validation was conducted with the testing dataset by means of Cross-Validation Folds by specifying the number of folds as 10. Efficiency of the model was measured by considering the accuracy. The results of the investigation on factors influencing dropout of students can be ordered according to their significance as follows:

**Table 4** Values of attributes that are responsible for dropout of students, ordered by their significance from the highest to the lowest

| Factors | Accuracy | Priority<br>(from highest to lowest) |
|---|---|---|
| Previous GP | 72.02 | 1 |
| Previous academic field | 70.11 | 2 |
| Previous education | 68.13 | 3 |
| Occupation of parents | 65.25 | 4 |
| Sex | 54.11 | 5 |
| Province of residence | 53.45 | 6 |

According to Table 4, it could be seen that the factors that have influence on dropout of students, as ranked by their significance from the highest to the lowest, were previous GPA, previous academic field, previous education, occupation of parents, sex, and province of residence. When taking these factors to examine their association by using the classification rules, it could be found that there were many causes affecting student dropout.

2.2 Association-rule analysis

During 2007-2015 academic years, students in the Department of Computer dropped out due to having a study result lower than the criteria show that there are 202 of them. These 202 students had their grades lower than C (namely D+, D, F, and E) for 2,568 records in total. These data were used to find association of courses that were most affect to poor study results. The association could be discovered with a technique to determine association rules. The Delta value was set to be equal to 0.05, the Low Bound Min Support was 0.1, and the minimum confidence (Minmetric) was 0.9. An example of the experiment is shown in Table 5.

**Table 5** An example of the experiment to determine association rules

| Association rules | Confidence | Support |
|---|---|---|
| GLAN11021==> COM 13041 | 1 | 0.24 |
| GLAN11021, GSCI11011==> COM 13041 | 1 | 0.22 |
| GSCI11011, MATH14011 ==> COM 13041 | 1 | 0.16 |
| PHYS1101 ==> COM 1301 | 1 | 0.14 |
| COM 1401, PHYS1102==> PHYS1101 | 1 | 0.14 |
| COM 26021 ==> COM 13041 | 1 | 0.11 |
| COM 1301, COM 1401, PHYS1102 ==> PHYS1101 | 0.97 | 0.14 |

According to Table 5, it was found that different courses may have some effect on each other, and eventually affect dropout of students. For example, GLAN 11021 (an English course) had some effect on COM 1304 (a Computer course), and PHYS1101 (a Physics course) has some effect on COM 1301 (a Computer course). In summary, Table 5 revealed that the courses being influential on dropout of the students are COM (Computer courses), GLAN (English courses), MATH and GSCI1101 (Math course), and PHYS1101 (Physics course).

## Discussions

From studying data of students in Chiang Mai Rajabhat University, it was found that many students had dropped out. Two important factors for dropout of students were their previous knowledge background and their current study results. Hence, this case study was conducted with students in the Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University. The students had 3 academic major namely Computer Science program, Information Technology program, and Programming and Web Security Maintenance program. In total, these 3 programs had 1,433 students during 2007-2015 academic years. Data of these students were divided into 2 sets. The first set contains 1,433 records that were used for studying factors related to previous education that influence dropout of the students. The analysis with this data set was done by using a data mining approach with a classification method based on the decision tree technique and C4.5 algorithm. The analysis results reveal that the significant factors on dropout of students as ranked by their importance from high to low are previous study result, previous study field, previous educational certificate, occupation of parents, and sex, respectively. The data about sex and province had only little impact on status of the students, by having accuracy values of only 54.11 and 53.45, respectively. The second data set was used to study factors of student dropout based on current study results. These data were collected only from the group of 202 students who had dropped out and also had learning grades of E, F, D, and D+ since these grades are considered to have impact on their dropout. These screening criteria yielded 2,568 records of data to be analyzed. The analysis was conducted with the association-rule data mining technique with Apriori algorithm in order to find association of courses taken by the students and led to their dropout. The experiment results reveal that the courses that had some effect on student dropout were Computer, English, Mathematics, and Physics. In addition, GLAN 11021 (an English course) was found to have some effect on COM 1304 (a Computer course), and PHYS1101 (a Physics course) had some effect on COM 1301 (a Computer course) etc. According to the results, the courses that have impact on dropout of the students are COM (Computer courses), GLAN (English courses), Math and GSCI1101 (Math course) and PHYS1101 (Physics course).

## Conclusion

The study to determine factors affecting dropout of students with the use of data mining technique based on decision tree classification and C4.5 algorithm found that the factors with high influence on student dropout include previous study result (GPA), previous study field, and previous certificate. That means these factors have impact on the current study results. This study also analyzed data of current study results with the association-rule technique, and found that the courses that affect the students' dropout were Computer, English, Math and Physics. During the period of 2007-2015 academic years, the university has improved its curriculum by emphasizing main courses and core courses of Faculty of Science and Technology. The curriculum's structure is emphasized on Mathematics, Sciences, and English. These changes are therefore influential toward study results of students.

## Recommendations

1. This study on factors of dropout was based only on the existing database of students. Factors beyond the database, such as students' behaviors, which may also impact their dropout, have not yet been studied.

2. There are also some attributes in the database that have not been included in this study, such as characteristics of the previous institution. Some characteristics such as school type (a public school, a private school, or a non-formal education school) and school location (being in an urban area or a suburb area) might also affect student dropout. Their influences deserve to be studied afterwards.

3. Most of variables taken from the database were mostly in Thai language. These Thai variables cause some problems when being used with Weka software. Thus it require quite much time and effort to adjust and check the data for maximum completion.

4. In order be able to make a more comprehensive conclusion, other methods based on the same classification data-mining technique should also be studied. Such these methods may include Naive Bayes and artificial neural network methods.

## References

Bean, J.P. (1980, June). Dropouts and turnover: The synthesis and test of a causal model of studentattrition. **Research in Higher Education, 12**(2), 155-187.

Chemers, M. M., Hu, L., & Garcia, B. F. (2001, March). Academic self-efficacy and first year college student performance and adjustment**. Journal of Educational Psychology, 93**(1), 55-64.

Hsieh, P., Sullivan, J. R., & Guerra, N. S. (2007, May). Closer look at college students : Selfefficacy and goal orientation**. Journal of Advanced Academics, 18**(3), 454-476.

Laokietkul, J. & Sitthiworachart, J. (2008). A Forecasting Model for Evaluate Freshmen's Quality in Information Technology Program with Class Association Rules. **Proceeding of the National Conference on Computer and Information Technology (NCCIT'08)**. 23-24 May 2008 (pp.701-706). Mahasarakham : rajabhat mahasarakham university.

Lee, D. H., Kang, S., & Yum, S. (2005, September). A qualitative assessment of personal and academic stressors among Korean college students*:* An exploratory study. **College Student Journal, 39**(3), 442-448.

Mahapatra, I. & Bose, R. K. (2001, December). Business Data Mining-a Machine Learning perspective. **Information and Management, 39**(3), 211-215.

Mohr, J. J., Eiche, K. D., & W.E. Sedlacek. (1998, July-August). So close, yet so far: Predictors of attrition in college seniors. **Journal of College Student Development, 39**(4), 343-54.

Pansumret, Y., Phuboon-ob, J., & Pongsiri, W. (2013, April). On Comparison of Data Mining Algorithms for Analysis of Factors Affecting the Academic Performance of Students. **Journal of Sciences, Mahasarakham University, 32**(Special Volume)**,** 281-289.

President's Offices, King Mongkut's University of Technology North Bangkok. (2011). **Report of student status 2011, King Mongkut's University of Technology North Bangkok.** 1.

Thongkon, W. (2013). **A Study of Factors affecting student Dropout of Huachiew Chalermprakiet University.** Office of Academic Development, Huachiew Chalermprakiet University. 119.

Suwannarattaphoom, P. & Waiyamai, K. (2012, January-March). An approach for improving associative classification in imbalanced datasets. **Kasetsart Engineering Journal, 25**(79), 36-49.

Werghi, N. & Kamoun, F. (2010). A decision-tree-based system for student academic advising and planning in information systems programmes. **Int. J. Business Information Systems, 5**(1), 1-18.

Willcoxsona, L., Julie, C. J., & Sally J. S. (2011, May). Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities. **Journal of Studies in Higher Education , 36**(3), 331-352.