Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out

WILAIRAT YATHONGCHAI¹, CHUSAK YATHONGCHAI¹, KITTISAK KERDPRASOP², NITTAYA KERDPRASOP²

¹School of Information Technology,  ²Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000, Thailand
y_wilairat@hotmail.com, y_chusak@yahoo.com, kerdpras@sut.ac.th, nittaya@sut.ac.th

Abstract:- The increase of students’ drop out rate in higher education is one of the important problems in most institutions. The discovery of hidden knowledge from the educational data system by the effective process of data mining technology to analyze factors affecting student drop out can lead to a better academic planning and management to reduce students drop out rate, as well as can inform valuable information for decision making of stake holder to improve the quality of higher educational system. In this paper, we consider three issues of factors affecting students’ drop out rate. These factors are conditions related to the students before admission, factors related to the students during the study periods in the university, and all factors including the target value to be predict for factors analysis. We use tree-based classification algorithm, J48 or C4.5, and Naïve Bayes to analyze the data. To evaluated the model, we use both 10-fold cross validation and supplied test set methods. Accuracy rate was satisfactory and the induced models are actionable and potentially applicable to higher education planning.

Key-Words:- Higher education, Student drop out, Data mining technique, Classification.

1. Introduction

Information technology has an important role in most organization that manipulates and collects data in large databases. Stored data can be used to generate useful information for decision making. Data mining is an automatic data analysis process that helps users and administrators to discover and extract patterns from stored data [1]. The use of data mining technique to analyze an educational database is absolutely expected to be of great benefit to the higher educational institutions.

Nowadays, the educational information such as the students’ information, the courses’ detail, the measurements and assessments, and so on, has increased tremendously. As a consequence, several factors have involved to affect the quality of higher educational system.

The quality is the major key performance factor in higher educational system. The acquirement of quality in higher educational system must be planned, monitored, and controlled in each and every education processes with the main purpose of improving the efficiency of students. The indicator of the quality weakness of the educational system is the large number of students that drop out. By the way, to predict the number of drop out students and factors affecting the drop out situation must use the effective processes. Dekker et al. [2] used data mining technique to predict the electrical engineering student drop out and identifying success-factors specific to the engineering students. Kotsiantis [3] also applied educational data mining techniques to predict drop out and school failure that is also important to resolve the problem. As well as M. Jadrić et al.[4] to analyze the problem of students drop out in the higher education by using the data mining methods and also a model suggested. There have been increasing research interest in the use of data mining in education development and discover knowledge from educational environments [5].

This paper aims to present the experience in using the educational information from knowledge base in an effective way by applying data mining techniques to analyze the major factors that affect
the drop out of students in the institutions of higher education. The main purpose of our study is to deploy the analysis results to improve the student learning ability, to decrease of the number of drop out students, and to convey actionable information that can facilitate decision making of teachers, education management team, or anyone involving in the teaching and learning system of the higher education.

2. Related Work

Data mining techniques have been successfully used to enhance various aspects of educational quality of higher educational system. Shaeeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, and M. Inayat Khan [6] used data mining technique named k-means clustering applied to analyze student’s learning behavior that will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students.

Sajadin Sembiring et al. [7] studied to apply the kernel method as data mining techniques to analyze the relationships between students’ behavioral and their success then they developed the model of student performance predictors which can help to predict the successful student by employing psychometric factors as variables predictors.

Xie Wu et al. [8] used data mining technique with data of undergraduates have to be stored in database or data warehouse with the capacity increasing. The method is carried out by decision tree algorithms. The results of case reveals that the decision tree algorithm of data mining technology can distinguish between the merits of the level of university students and realize the classification comprehensive evaluation, and solve the problem that the traditional methods are not fit for the student assessment of too much records, which greater efficiency.

Diego Garcia-Saiz [9] compared the performance and interpretation level of the output of the different classification techniques to applied on case study from a course offered in the last three academic years (2007-2010) at the University of Cantabria and propose a meta-algorithm to pre process the datasets then improve the accuracy of the model.

J.F. Superby et al. [10] to classify 533 first-year university students into three groups: the low-risk, the medium-risk, and the high-risk students(high probability of dropping out) and provides the most significant variables correlated to academic success. They are gathering data on November of academic year 2003-04. The result of the application of data mining methods to predicting students’ academic success.

Al-Radaideh et al [11] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. This research used three classification methods namely ID3, C4.5, and the Naïve Bayes. The results indicated that Decision Tree model had better prediction than other models.

The work appeared in [12][13] used data mining techniques to increase the efficiency in higher educational system by focusing on the academic performance, evaluation and classification of students for decision making to evaluate the quality of students.

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making.

3. Research Methodology

Information produced by data mining techniques can be represented in many different ways. In this paper we have used the classification data mining technique to extract the important attribute that stored in a database to analyze factors affecting the drop out of students in higher education by two classifiers algorithms, J48 and Naïve Bayes.

3.1 Classification

Classification technique is usually use in data mining which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In learning, the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples.

Decision tree structures are a common way to organize classification schemes. In classifying tasks, decision trees visualize what steps are taken to arrive at a classification. Decision trees are the classic way to represent information from a machine learning algorithm, and offer a fast and powerful
way to express structures in data. The J48 algorithm gives several options related to tree pruning. Pruning produces fewer, more easily interpreted results. The basic algorithm described above recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible. This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular idiosyncrasies of that data. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy [14].

Naïve Bayes is one of the most effective and efficient classification algorithms. This classifier is based on the Bayes’ Theorem and the maximum posteriori hypothesis. The naive assumption of class conditional independence is often made to reduce the computational cost [15].

3.2 Study Framework
The study framework includes 4 steps: Data preprocessing, Classifier algorithm, Evaluation classifier and Build academic DSS model as shown in Fig. 1.

4. Data mining Process
The challenges in data mining are scaling the algorithms to work with large datasets and format variety of data. The data extracted from database of educational information is dynamic and difficult for the experimentation phase.

The data warehouse at Buriram Rajabhat University stored databases for working in the university such as academic database, financial database, quality assurance database and so on, which useful for the administration of the university.

4.1 Data Preparation
The data set used in this study was obtained from the database of Academic MIS at Buriram Rajabhat University (BRU) in Thailand between 2008 and 2009. Sample data were from faculty of science which has the highest students drop out rate in this university. There are 731 students enrolled in bachelor degree. All of the students contain 481 students during the study periods in the university and 251 drop out students. In this step, data stored in different tables was joined in a single table. After joining process, errors were removed.

4.2 Data selection and transformation
In this step, only fields were selected which were required for data mining. While some of the information for the attribute was extracted from the database and transform values for data mining.
In this paper, we defined the assumption of the factors affecting students’ drop out rate are factors related to the students before admission and factors related to the students during the study periods in the university. Therefore, the analysis points have three issues.

- Factors related to the student before admission - The student background that effect to student drop out such as GP AX from high school, program to study from high school and school size.

- Factors related to students during the study periods in the university – These factors are the major causes of students to drop out. The factors include program of study, GPA score from the first 4 terms, and student loan.

- All factors - All of above including cause to drop out, drop term and drop out status are the target value to be predicted for factors analysis that effect to student drop out.

All variables in this experiment are shown in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>Program to study in faculty of science</td>
<td>[230, 240, 241, 243, 247, 249, 264, 265, 284, 285, 286]</td>
</tr>
</tbody>
</table>
| GPA1-GPA4        | GPA in Term1-Term4 (in Academic year 2008-2009)  | [Weak, Medium, Good, Best] →  
                    |                                                 | Weak = GPA < 1.6 
                    |                                                 | Medium = GPA 1.6-1.99 
                    |                                                 | Good = GPA 2.0-2.5 
                    |                                                 | Best = GPA > 2.5 |
| SchoolGPAX       | GPAX from high school                            | number                           |
| SchoolProgram    | Program to study in high school                 | {1, 2, 3} →  
                    |                                                 | 1 = science + math 
                    |                                                 | 2 = language + math 
                    |                                                 | 3 = other |
| SchoolSize       | Size of school                                  | [Small, Medium, Large]           |
| Loan             | Student loan                                    | {Yes, No} →  
                    |                                                 | Yes = has a student loan 
                    |                                                 | No = has not a student loan |
| Cause            | Cause to drop out                               | {Studying, Retired, Finance, ChangeU, ChangeProgram, Relocated, Stop} |
| DropTerm         | Term to drop out                                | {1, 2, 3, 4, 5, 6, No}           |
| DropOut          | Drop out status                                 | {Yes, No}                        |

5. Analysis Results

According to the experimentation, we analyzed factors affecting student drop out in higher education by comparing the two classifier algorithms which are J48, or C4.5, and Naïve Bayes. Weka 3.6.5 was used as the tool for this study.

The results of this research (the analysis points have three issues) in the form of decision tree training modeling; therefore we require interpreting and generating explanation, which is understandable by humanity. Therefore the obtained decision tree is translated into rules. We would describe the interested rules as follow:

- The student who has a student loan will not drop out while who has not will drop out if GPAX from high school less than 2.42.
- The student who has not a student loan and GPAX from high school more than 2.42 and studied program in high school was Science-Mathematics will not drop out while who studied other program will drop out.
- The student who drops out after finish first term would have first term GPA less than 1.6 and has not a student loan.
- The student who has first term GPA less than 1.6 and has a student loan will drop out after finish second term.
- The student who has first term GPA: 1.6-1.99 and second term GPA less than 1.6 will drop out after finish second term.
- The student who has fourth term GPA more than 1.6 will not drop out.
- The student who studied in Sports science (major ID=240) will not drop out.
- The student who studied in Community health or Computer science (major ID=265 or 230) will have high drops out rate during first term as a sequent.
- The student who studied in Information technology or Computer Technology (major ID=284 or 286) and has first term GPA more than 2.5 but second and third term GPA less than 1.6 will drop out after finish fourth term.
- Most of students who drop out during first term because they want to change major and will re-entrance in the next year while some of them have a finance problem, relocated, change university and have no reason.

We can use the rules to improve student admission plan, tracking and help the students who have a high probability of dropping out including educational quality management planning of the university.
The supplied testing set and 10-fold cross validation are the methods that we used to evaluate the model. In supplied testing set method, all data were split into two parts (training set and testing set). By 30% of instances from each program in faculty of science were random separated to testing set (218 instances) and data remaining were training set (513 instances).

The data analysis by using before admission factors that affecting the students drop out aims to analyze the characteristic of the student who want to study in science faculty. The result values of evaluation are shown in table 2.

Table 2 Comparison of results of two classifier algorithms on before admission factors.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>J48</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validation set</td>
<td>Testing set</td>
</tr>
<tr>
<td>Accuracy</td>
<td>78.39%</td>
<td>76.60%</td>
</tr>
<tr>
<td>TP Rate</td>
<td>0.784</td>
<td>0.766</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.312</td>
<td>0.371</td>
</tr>
<tr>
<td>TN Rate</td>
<td>0.779</td>
<td>0.756</td>
</tr>
<tr>
<td>FN Rate</td>
<td>0.784</td>
<td>0.766</td>
</tr>
</tbody>
</table>

The data analysis by using students during the study period in the university factors aims to know how the student’s grades affect to the student drop out. The result values of evaluation are shown in table 3.

Table 3 Comparison of results of two classifier algorithms on studying student factors.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>J48</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validation set</td>
<td>Testing set</td>
</tr>
<tr>
<td>Accuracy</td>
<td>87.14%</td>
<td>85.78%</td>
</tr>
<tr>
<td>TP Rate</td>
<td>0.871</td>
<td>0.858</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.068</td>
<td>0.065</td>
</tr>
<tr>
<td>TN Rate</td>
<td>0.843</td>
<td>0.845</td>
</tr>
<tr>
<td>FN Rate</td>
<td>0.871</td>
<td>0.858</td>
</tr>
</tbody>
</table>

The data analysis by using all factors aims to know what the factors affect to the student drop out. The result values of evaluation are shown in table 4.

Table 4 Comparison of results of two classifier algorithms on all factors.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>J48</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validation set</td>
<td>Testing set</td>
</tr>
<tr>
<td>Accuracy</td>
<td>87.00%</td>
<td>84.86%</td>
</tr>
<tr>
<td>TP Rate</td>
<td>0.87</td>
<td>0.849</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.073</td>
<td>0.066</td>
</tr>
<tr>
<td>TN Rate</td>
<td>0.843</td>
<td>0.831</td>
</tr>
<tr>
<td>FN Rate</td>
<td>0.851</td>
<td>0.849</td>
</tr>
</tbody>
</table>

Comparison of accuracy of two classifier algorithms from table 2-4 are shown in Fig. 2.

Fig. 2 Comparison of accuracy chart

The accuracy of two classifier was found to be no different. And within an acceptable level.

6. Conclusion and Future work

Factors Analysis in Higher Educational Student’s Drop Out is an important. In this paper we presented the effectiveness of classification techniques (J48 and Naïve Bayes algorithms) on the data set used from the database of Academic MIS at BRU. Sample data were faculty of science. The three issues of factors analysis affecting to student drop out are: factors related to the student before admission, factors related to the students during the study periods in the university, and all factors. Our experimental results are shown as the rules that transformed from decision tree by accuracy value between 75% and 88%. Based on the three issues analysis, we found the fundamental factors about...
student before admission to planning to qualify for admission. The knowledge about students during the study periods in the university factors can use for academic planning to improve the quality of students.

**Suggestions**

1. Data preparation in data mining process is very important. The experience from this research, we have been used data stored in educational database to analyze which there are many tables, several format and large number of records. We need to merge data. So, we required good planning and provided the data preparatory steps are carried out carefully.

2. Attributes selection for factors analysis affecting students drop out is very important to data mining processes. Appropriate attributes for data classification, we found that the data values should be repeated and not various.

3. From the partial results of research, students’ drop out rate in the first year of the student is more highly than other years. Therefore, the higher educational system should be given priority to the new student in both academic and behavior.

After we have the knowledge about factors affecting student drop out. Our future work is using data mining technique to evaluate performance of students in higher education to improve the better quality of education.

**References**


