An Improved Recommendation Models on Grade Point Average Prediction and Postgraduate Identification using Data Mining

KANOKWAN WATKINS
Didyasarin International College,
Hatayi University, Hatayi, Sogkhla 90110
THAILAND
kokoya002@yahoo.com, kanokwan@hu.ac.th http://www.hu.ac.th

Abstract: - Numerous educational institutes have established technical support services intending for greater completion rates in tertiary education. One form of services providing by all universities is student counseling. In order to assist supervisors and counselors, this study aims to apply techniques and methodologies to develop a model based on classification techniques using past cases from student database to predict the likely student’s Grade Point Average (GPA) of prospective student and current students. Also, predict final year students or graduates to identify potential students who may continue with postgraduate study. In the experiment, two datasets were used. The results are interpretation of the performance of the new models based on ANN, SVM, CHAID, Ensemble and MANN-OWSR techniques. The experimental results found that the proposed model enhances the accuracy of Data Mining techniques in comparison to the benchmark model.

Key-Words: - Counselling, recommendation system, data mining, CHAID algorithm, SVM, ANN

1 Introduction

The number of successful completions of students within an educational institute is part of any performance criteria, however not all students that enrol will complete successfully as records have indicated [1]. The failures can be attributed to high rates of drop-outs from the courses or other factors faced by the students. This is highly undesirable as this means wasted resources and reduced number of graduates to meet the demands by the industry and the community. There are many causes to this problem, and they could occur anytime during the course of study. One of the factors may be due to the enrolment process where a student has not selected or has not received recommendations for courses that are more suitable for the student.

With limited resources and increasing competition for students in the education sector, higher education institutes are focusing on efforts to increase the rate of student retention and completion. In addition, a university’s performance is also increasingly being used to measure its quality and “reputation” [2]. One aspect of such measurement is based on factors which affect a student’s satisfaction. For instance, Gatfield [3] has concluded that it is significant for higher education institutes to concentrate on the issue of quality, through accreditation processes and various aspects of quality services from the students’ perspectives. Another aspect of student services is counselling.

Archer and Cooper [4] confirmed that provision of counselling services is an important factor contributing to students’ academic success. In addition, Urata and Takano [5] stated that the essence of the student counselling should include advices on career guidance, identification of learning strategies, handling of inter-personal relation, along with self understanding of the mind and body. It can be said that a key aspect of student services is to provide counselling on course guidance as this will assist the students in their decision and future university experience. Many students choose particular courses of study just because of perceived job opportunities. However, issues may arise if a student is not interested in the career, or if the course is not suitably matched with the student’s capability [6]. In the tertiary education sector, teaching staff may have insufficient time to counsel the students due to workload and there are inadequate tools to support them. Hence, it is desirable that some forms of intelligent recommendation tools could be developed to assist students in or before they enrol in the universities. This forms the motivation of this study.

This study aims to develop and apply techniques and methodologies based on classification techniques using past cases from the student database to predict the likely GPA results of prospective, new and current students. The aim is to assist supervisors and counsellors to advise...
prospective and enrolled students, and to predict the likely results from postgraduate study to identify potential students who may continue with postgraduate study, and to improve the performance of the recommendation model by using combination techniques, that is, the ensemble and MANN-OWSR methods, finally to propose the techniques to be used in the model and choose the best model with the highest accuracy for use in the intelligent recommendation system.

This paper is separated into various sections. The next section explains the student counselling in universities followed by justification for the proposed recommendation models. Section 4 illustrates a discussion of the techniques used. Section 5 presents the input and output variables selection, including an explanation of the datasets. The experimental design and experimental results followed in Section 6 and 7, respectively. The final section explains the conclusion and discussion of the study.

2 Student counselling in universities
One type of service that supports SRM and is provided by most universities is student counselling. Archer and Cooper [7] stated that the provision of counselling services is an important factor contributing to students’ academic success. Further, the advancement of technology in educational institutions could create opportunities for substantial improvement in management and information systems. Many designs and techniques now allow for better results in analysis and recommendations. With this in mind, universities in Thailand are working towards improving education quality [8] and many institutes are focusing on how to increase student retention rates and completion rates. In addition, a university’s performance is also increasingly being used to measure its ranking and reputation [9]. Urata and Takano [10] stated that the essence of student counselling should include advice on career guidance, identification of learning strategies, handling of interpersonal relationships and self-understanding of the mind and body. It can be said that a key aspect of student services is to provide programme guidance, as this will assist the students in their programme selection and future university experience. Other research focused on the provision of counselling and careers services, which have been adopted by many universities. To enhance the university’s mission, the prominent services provided by universities are psychological counselling, careers and work-placement advice and financial assistance.

Conversely, many students have chosen particular programmes of study because of perceived job opportunities, peer pressure and parental or family advice. Issues may arise if a student is not interested in the programme or if the programme or career is not suitably matched with the student’s capabilities [11]. In Thailand’s tertiary education sector, teaching staff may have insufficient time to counsel students because of high workload and inadequate support tools. Hence, it is desirable that some form of intelligent recommendation tool was developed to assist staff and students in the enrolment process. This forms the motivation of this research.

3 Justification for the proposed recommendation models
Prior studies have addressed issues faced by Thai students during their time at university. For example, Sarawut [12] studied the causes of dropouts and programme incompletion among undergraduate students from the Faculty of Engineering at King Mongkut’s University of Technology North Bangkok. It was reported that the general reasons for under achievement were due to teaching and learning issues. Further, the study showed that there were three groups, which each had different reasons for not completing their studies. The first group’s primary reason for incompletion was the students’ attitude towards the field of study. This group felt that their field of study was too difficult. The second and third group’s primary reasons were related to teaching and learning. Hence, this indicated the need to match the programme requirements with the academic capabilities of the students.

Another study at the Dhurakij Pundit University, Thailand, examined the relationship between learning behaviour and low academic achievement (below 2.0 GPA) of first-year students in regular four-year undergraduate degree programmes. The results indicated that students who had low academic achievement had a moderate score in every aspect of learning behaviour. On average, the students scored the highest in class attendance, followed by the attempt to spend more time on study after obtaining low examination grades. Some of the problems and difficulties that affected students’ low academic achievement were students’ lack of understanding of the subject and the lack of motivation and enthusiasm to learn [13].
While most Thai students considered a university degree an essential part of their education, many of them did not know which programme and subjects to study. One service that can help students and staff with this challenge is the student counselling service, which provides programme advice and counselling for new students to achieve a better match between the student’s ability and the chances of success in completing the programme. In private universities in Thailand, this service is normally provided by counsellors or advisors who have many years of experience in the organisation or in Higher Education. However, with the increasing number of students and expanding number of choices, the workload on advisors is becoming too much. It is apparent that some form of intelligent system will be useful in assisting the advisors and this forms the motivation of this study.

In summary, it is necessary to meet student needs and to match their capability with the programme of their choice in the recruitment and enrolment of students in Thai universities. The students’ backgrounds may also have a part to play in the matching process. Understanding student needs will implicitly enhance the student’s learning experience and increase their chances of success, thereby, reducing resource wastage that is due to dropouts and change of programs. Therefore, these factors are considered in the proposed recommendation system in this study. The techniques used are described in the following section.

4 Intelligent Techniques

The first classification technique chosen for the GPA recommendation model was SVM. Many research reports have shown that SVM is capable of providing successful outcomes from classification tasks [14, 15, 16].

The second technique used in this framework is ANN, which is a data-driven, self-adaptive method and is also a successful and popular technique in classification [17, 18, 19, 20, 21, 22]. The third classification technique is the DT algorithm, which has been used in various studies [23, 24, 25], as well as in many educational data-mining studies. In this study, the DT algorithm based on CHAID was used [26, 27]. Ramaswami and Bhaskaran [28] reported that the results from the CHAID algorithms were satisfactory and the CHAID algorithms could also be used to analyse both binary and categorical data. As many feature variables in this study were categorical data, the DT based on CHAID algorithm was deemed an appropriate technique for the intelligent recommendation system.

In general, combined classification models can improve the prediction performance of the classifiers [29, 30]. In this study, two main aggregation techniques were employed. One was the ensemble method based on confidence-weighted voting. One study showed that ensemble is able to reduce prediction errors; however, it depends on the model variance of the classifiers [30]. The other aggregation technique used in the study was the MANN-OWSR, which is an efficient aggregation technique introduced and reported by Kajornrit [31]. This technique has shown an acceptable improvement accuracy, and it can be used to combine two classification models [31]. This suits the methodological design in this study and the input and output variables selection is described in the following section.

5 Input and Output Variables Selection

In this study, two sets of data were organised during the pre-processing data stage. In the two likely GPA modules, the datasets used were the same. The sample data of these two modules were chosen from the university’s database of 11,400 student records. After the data cleaning process, 9,001 student records were used in this study. The distribution of the students, with respect to programmes, is illustrated in the following figures.

Fig 1. The number of undergraduate students in each course of study (2001-2007)

In Fig 1, the tertiary student data were obtained from seven academic years of records (2001–2007), excluding summer semesters. Student data included records from first year to graduation. The data comprised of 30.62 per cent of students from business computing, 19.02 per cent from accounting, 22.18 per cent from management, 14.75 per cent from marketing, 5.2 per cent from human
resource management, 4.84 per cent from business English and 3.38 per cent from law. The data in this study did not indicate any personal information because of privacy issues, and no student was identified in the research. The university randomised the data and all private information was removed in this experiment.

The postgraduate study identification module after the data cleaning process are illustrated below.

The postgraduate study identification module after the data cleaning process are illustrated below.

![Graph](image)

**Fig 2:** Number of postgraduate students in each postgraduate programme (2001–2009)

Fig 2 shows that the dataset of 918 postgraduate student records from nine academic years (2001–2009, excluding summer semesters) is made up of 38 per cent of students from the Master of Education in Educational Administration, 36 per cent from the Master of Business Administration, 16 per cent from the Master of Education in Curriculum and Instruction and 10 per cent from the Graduate Diploma in Teaching Profession.

In this study, the process of choosing variables was based on results from a survey conducted and provided by the university. The variables used in these two modules are shown in Table 1.

The selection of appropriate input feature variables is essential for classifiers. In this study, previous school GPA, interests and gender are associated with the ability of students to study at the tertiary level [32-34, 27]. Therefore, the main variables chosen in Module 1 and 2 were previous school GPA, talents and interests and gender. However, other parameters may be useful in data analysis by data mining, and additional supportive variables used in this experiment are shown in Table 1. The module for identifying students who are likely to succeed in postgraduate study used similar variables to the other modules in this study. As choosing an appropriate activity could also improve student performance at university [35], this variable was also used in Module 3.

<table>
<thead>
<tr>
<th>Table 1. Variable names and data types in two likely GPA modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 1: likely overall GPA for new students</td>
</tr>
<tr>
<td>Variable name</td>
</tr>
<tr>
<td>Overall GPA</td>
</tr>
<tr>
<td>Pre-GPA</td>
</tr>
<tr>
<td>Pre-major</td>
</tr>
<tr>
<td>Type of school</td>
</tr>
<tr>
<td>No. of Pre-awards</td>
</tr>
<tr>
<td>Talents and interests</td>
</tr>
<tr>
<td>Motivation channels</td>
</tr>
<tr>
<td>Admission round</td>
</tr>
<tr>
<td>Guardian occupation</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Major</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Variable names and data types in the module of Identification of potential students to continue with postgraduate study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 3: Identification of potential students to continue with postgraduate study</td>
</tr>
<tr>
<td>Variable name</td>
</tr>
<tr>
<td>Master degree success</td>
</tr>
<tr>
<td>University GPA</td>
</tr>
<tr>
<td>Postgraduate major</td>
</tr>
<tr>
<td>University major</td>
</tr>
<tr>
<td>University awards</td>
</tr>
<tr>
<td>Type of university</td>
</tr>
<tr>
<td>Pre-school GPA</td>
</tr>
<tr>
<td>Type of school</td>
</tr>
<tr>
<td>Motivation channels</td>
</tr>
<tr>
<td>Guardian occupation</td>
</tr>
<tr>
<td>Activity type</td>
</tr>
<tr>
<td>Gender</td>
</tr>
</tbody>
</table>

The next section discusses the intelligent techniques used.
6 Experimental Methodology and Design

Fig 3 illustrates the process that determines the best GPA recommendation model for use in this study. While the three techniques (ANN, SVM and CHAID) have been used extensively in the past, it is recognised that the ensemble and MANN-OWSR methods have the potential to improve accuracy. Hence, it is necessary to determine whether the single or combined model should be used. The process adopted in this study is described below.

First, the ANN, DT based on CHAID algorithms and SVM were used. The results from these three models were then compared in the first result comparison.

Second, the two models that returned the lowest performance accuracy were combined using the ensemble approach based on the confidence-weighted voting method. Then, the result from the ensemble model was compared with the results from the three models. The two models that gave the best results were chosen for the next process.

Next, the two models with the highest accuracy from the previous comparison were aggregated using MANN-OWSR. Then, the model (SVM, ANN or CHAID) with the best result was compared with the results from the ensemble and MANN-OWSR models. The one that returned the best performance accuracy and the lowest accuracy error rate was then chosen to predict the overall GPA for prospective, new and current students.

This technique was also applied to determine the best model for the prediction of results from postgraduate study to identify potential students to continue with postgraduate study. After the model was determined, it could be used by counsellors and supervisors to provide recommendations for the students.

The outputs from the likely overall GPA and GPA for each semester module were categorised into six GPA classes; for example, A is likely to get a GPA of 0.3, which is between 2.254 and 2.720 (see Table 3). Outputs from the postgraduate identification module were provided in five postgraduate GPA categories. For example, if B is a senior student and the result shows that B is likely to obtain a postgraduate GPA of 0.3, this refers to the GPA range 3.4–3.59. Examples of the results are given in Table 3 and 4.

Table 3. Example results for likely overall GPA and likely GPA in each semester

<table>
<thead>
<tr>
<th>Student no.</th>
<th>Likely GPA</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A001</td>
<td>2.254–2.720</td>
<td>Performance of this student needs to be monitored and counselling should be provided, if needed</td>
</tr>
</tbody>
</table>

Table 4. Example results for postgraduate identification

<table>
<thead>
<tr>
<th>Student no.</th>
<th>GPA class</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>B009</td>
<td>3.4–3.59</td>
<td>This student is likely to be successful in postgraduate study with good results</td>
</tr>
</tbody>
</table>
7 Experiment Results
This section provides example results from the GPA recommendation model. Three modules were developed: likely overall GPA (4Y), likely GPA of each semester (Y1S1 to Y4S2) and postgraduate identification (PG). To determine the best model, the experiment was trained, validated and tested three times to ensure consistency of results. The comparison results shown in this experiment are the average accuracy rate and MAE from each technique used in the model and the statistical probability of occurrence of the compared techniques. As described previously, SVM, ANN and CHAID are used in the first process.

7.1 First comparison between SVM, ANN and CHAID
The comparison results are illustrated in Fig 4.

In Fig 4, the accuracy rate between ANN and CHAID are similar: ANN performed slightly better in the likely overall GPA and GPA each semester module, but CHAID performed better in the postgraduate identification module. However, SVM performed considerably higher than ANN and CHAID in overall GPA and GPA each semester, while performing a little higher in postgraduate identification. To consider the MAE results in Fig 4, the comparison showed that the trend is similar to the accuracy rate.

Fig 4. Accuracy rate of the classification techniques

The comparison results of the first process demonstrated that SVM outperformed the ANN and CHAID techniques for all modules. In likely overall GPA and GPA each semester module, the second accuracy model was ANN, followed by CHAID as the lowest accuracy model. Conversely, in postgraduate identification, the second and third accuracy models were CHAID and ANN, respectively.

Therefore, the SVM-based model could be used to predict students’ GPA results with the greatest degree of accuracy in the first process. As SVM demonstrated the highest accuracy, it was considered in the second result comparison. However, as ANN and CHAID ranked second and third in accuracy for likely overall GPA and GPA each semester, they were combined by ensemble in the next process, while CHAID and ANN, which were ranked second and third in accuracy for postgraduate identification, were combined by ensemble in the next process for that module.

7.2 Second comparison of the ANN, CHAID and ensemble models
In order to improve the two lowest performing models, CHAID and ANN were combined by ensemble. The comparison results are shown in Fig 5.

Fig 5. Comparison of the accuracy rate between ANN, CHAID and ensemble

In the likely overall GPA and GPA of each semester module, the comparison results above show that the ensemble of the ANN and CHAID models slightly outperformed the individual ANN model, which was the second highest accurate model in the first process. In addition, the results of the average accuracy and MAE presented a similar trend: likely overall GPA (4Y) scored the lowest MAE and highest accuracy, whereas likely GPA in the first semester of Year 1 scored the highest MAE and lowest accuracy. Having considered most cases, the results of the ensemble model returned higher performance accuracy than the individually trained models; however, the ensemble method generated relatively small improvement in performance accuracy. In the postgraduate identification module, the results of the ensemble of the CHAID and ANN models showed slightly lower performance than the individual CHAID model. Further, having considered the above graphs and tables, the results indicated that the average performance of the combined ANN and CHAID models by ensemble outperformed the individual ANN and CHAID models in the likely overall GPA and GPA of each
year module. Therefore, in the next step, ensemble was chosen to combine with the SVM model, which was the highest accuracy model from the first process, using MANN-OWSR. Conversely, in postgraduate identification, CHAID showed higher performance than ensemble; therefore, CHAID was chosen to combine with SVM, also using MANN-OWSR, in the next process.

7.3 Third comparison using MANN-OWSR, SVM and ensemble in overall GPA and GPA of each semester
In these two modules, ensemble was combined with SVM in the aggregation techniques using MANN-OWSR. The comparison results are shown in Fig 6 and 7.

![Fig 6. Comparison of the accuracy between MANN-OWSR, SVM and ensemble](image1)

![Fig 7. Comparison of MAE between MANN-OWSR, SVM and ensemble](image2)

The figures above show that MANN-OWSR provided better accuracy and less prediction errors than ensemble but returned slightly better accuracy and less prediction error than SVM. The MAE comparison results also showed similar trends between SVM and OWSR. Considering the accuracy results together with MAE, the average performance of MANN-OWSR outperformed the individual SVM and ensemble models in these two modules. In the next section, the third comparison results of the postgraduate identification module are demonstrated.

7.4 Third comparison of MANN-OWSR, SVM and CHAID in the postgraduate identification module
In this module, the CHAID and SVM models were combined using MANN-OWSR. The comparison results are shown in Fig 8 and 9.

![Fig 8. Comparison of the accuracy between MANN-OWSR, SVM and CHAID](image3)

![Fig 9. Comparison of MAE between MANN-OWSR, SVM and CHAID](image4)

In the postgraduate identification module, the results showed that MANN-OWSR and SVM returned similar accuracy and both models returned higher accuracy than CHAID. The MAE comparison results showed similar trends to the accuracy results: MANN-OWSR and SVM returned the same results at 0.017 and CHAID had more errors than the first two models at 0.011. Even though the results of SVM and MANN-OWSR were similar, the average performance of SVM outperformed MANN-OWSR and CHAID. The results of the SVM model can be used to predict the best results of the postgraduate identification module model with the best degree of accuracy.

8 Conclusion and Discussion
This study proposed a process to develop the GPA recommendation model, which forms the three modules in the intelligent recommendation system. The first two modules focused on predicting the likely overall GPA for prospective and new students and the likely GPA for students in each academic year. The postgraduate identification module focused on final year students who were likely to be
successful in postgraduate study, and the result from this module could be used to support the scholarship committee and university administrator to estimate the number of potential students to carry on with postgraduate study.

This study also showed that the SVM model outperformed the ANN and CHAID models in the first process. The finding also indicated that the best recommendation model for the likely overall GPA and GPA in each semester module was the MANN-OWSR model. Conversely, the best model for the postgraduate identification module was the SVM model. This study demonstrated the use of intelligent techniques to determine the best model for predicting students’ GPA and identifying their potential to continue with postgraduate study. However, it is noted that datasets from other universities may exhibit different characteristics and the best model to be used may not be the same as in this study. The proposed model and process in Fig 3 provided an innovative approach to determining the best model for the prediction.

The next study discusses the identification of dropouts so that appropriate remedial actions can be initiated by the university to improve the retention rate.

References:


[25] Q.A. Al-Radaideh et al., 'Mining student data using decision trees', in International Arab Conference on Information Technology (ACIT2006), Yarmouk University, Jordan.


[27] N. Jantarasap. 'The study of the relationship between learning behavior and low academic achievement of Dhurakij Pundit University students' [Online].


