Using Association Rule Mining to Find the Effect of Course Selection on Academic Performance in Computer Science I

Lebogang Mashiloane

School of Computer Science,University of the Witwatersrand,
Johannesburg, South Africa
lebogang.mashiloane@wits.ac.za

Abstract. It is important for first year students in higher educational institutions to get the best advice and information with regards to course selection and registration. During registration students select the courses and number of courses they would like to enroll into. The decisions made during registration are done with the assistance of academics and course coordinators. This study focuses on the first year Computer Science students and their overall academic performance in first year. Computer Science I has Mathematics as a compulsory co-requisite, therefore after selecting Computer Science I, the students have to enroll into Mathematics and then select two additional courses. Can data mining techniques assist in identifying the additional courses that will yield towards the best academic performance? Using a modified version of the CRISP-DM methodology this work applies an Association Rule Mining algorithm to first year Computer Science data from 2006 to 2012. The Apriori algorithm from the WEKA toolkit was used. This algorithm was used to select the best course combinations with Computer Science I and Mathematics I. The results showed a good relationship between Computer Science I and Biology on its own, Biology with Chemistry and Psychology with Economics. Most of the rules that were produced had good accuracy results as well. These results are consistent in related literature with areas such as Bio-informatics combining Biology and Computer Science.

Keywords: Educational Data Mining, Association Rule Mining, Apriori algorithm.

1 Introduction

Registration is one of the most important days in a first year students' academic career. On this day, the student makes decisions about which courses and how many courses they would like to enroll into. These decisions could potentially influence that students' academic performance. At Wits, registration is done at the beginning of the year. Each group of students (by faculty and year of study) get a day or more devoted to them for registration. During registration

R. Prasath et al. (Eds.): MIKE 2014, LNAI 8891, pp. 323–332, 2014.

course coordinators and academics advise and assist the students. Some students are prepared and have done adequate research on the available courses and the courses' prerequisites. While other students have little knowledge of the available courses and the details around registration.

The advice that is given by the course coordinators and academics, present at registration, are based on co-requisites and/or past experience. This makes the advice subjective. Is there a way that the experience of these academics could be supported by results from an investigation of the historical data? Can data mining techniques be used to extract knowledge that could assist with this course selection and influence the best possible outcome? Focusing more specifically at the first year Computer Science students, how do the additional courses they select influence overall performance in first year? Data mining techniques will be used to attempt to identify the courses that, when combined with Computer Science, yield the best possible academic performance for first year students. This study will try to answer the following question:

How strong is the association between the courses selected during registration by Computer Science I students and the overall academic performance?

2 Background

2.1 Computer Science I Class and Registration

All first year students in the Faculty of Science at Wits are required to select atleast four courses to register into. Most courses have recommended and/or compulsory co-requisite courses. Computer Science I has Mathematics I as a compulsory co-requisite and Computational and Applied Mathematics (CAM) as a recommended co-requisite. Therefore all Computer Science I(`CS-I`) students have to register into Mathematics I. Additionally, most `CS-I` students register into CAM. There are many courses available for the `CS-I` students to select from. Science students are allowed to select courses from most of the schools in the university including schools from other faculties. The registration period is during the first month of the academic year and all the students in each year of study are given dates to come register. During registration lecturers and administrators from the different schools are gathered in a hall to advise and register the students into their selected courses. Some courses have pre-requisites and/or requirements of minimum marks for specific high school subjects. This information is available in the Wits prospectus book and provided during the registration process.

Apart from the courses that the `CS-I` students select, within the Computer Science course itself, these students are required to complete four modules. These modules are namely: Basic Computer Organization (BCO), Data and Data Structures (DDS), Fundamental Algorithmic Concepts (FAC) and Limits of Computation (LOC).

2.2 Association Rule Mining

Association Rule Mining is a data mining technique also known as Relationship Mining. The purpose of this technique is to extract frequent patterns or associations in a database or data set [1]. These patterns come in the form of a rule. The rules consist of an antecedent implying a consequent [1]. Association Rule Mining can be used to find frequent or least frequent item sets [7], this study will look for frequent item sets.

The Apriori algorithm is the most commonly used of the available Association Rule Mining algorithms. A high level description of the steps is shown in Figure 2.2. This study uses the predictive Apriori algorithm from the WEKA toolkit.

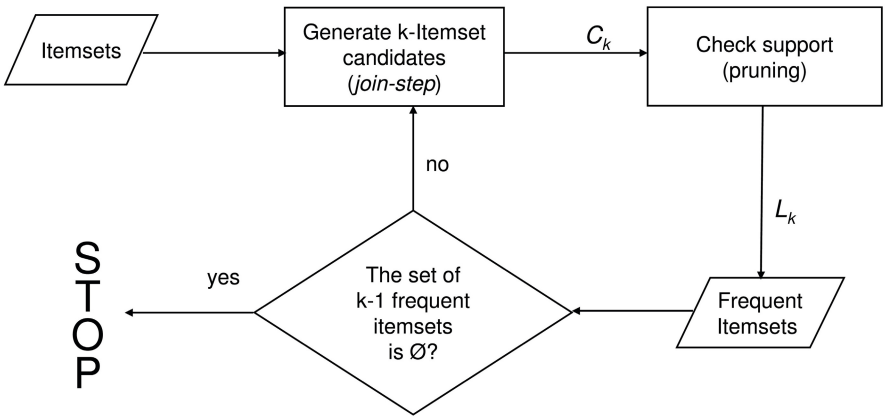

Fig. 1. Apriori Algorithm [5]

2.3 Related Work

Related work has shown ways in which data mining can be used to assist in recommending courses to students. A study by [4] aimed to create a tool that recommends courses to students. The four Apriori algorithms from the WEKA toolkit were used on a dataset which consisted of records from the Department of Information Technology, Computer Science and Engineering. The study found that students interested in Switching Theory & Logic Design and Operating Systems are more inclined towards Data Structures. The recommendation system suggested in this research focuses on suggesting courses based on profiling, therefore using the idea that students with similar profiles will prefer the same courses.

Another study by [6] analyzed data from 2002 to 2008 from the School of System Engineering at Universidad de Lima. This university uses an online registration system, therefore the students can enroll in some or all of the courses in their study plan or curriculum without consultation with any academics. [6] uses classification to create decision trees that would produce rules based on

the selection of courses of previous students and their academic performance. The approach taken was using the data to predict the best academic outcome by looking at the course selection using the C4.5 algorithm. Four iterations of the algorithm were done. Conclusions in the paper showed that satisfactory results can be found when trying to use data mining to recommend the number of courses and which courses to take for students. Another benefit that was found is that it also assisted with improving academic performance of the students.

[8] conducted a study at Griffith University, involving 251 students enrolled in courses such as drama, music, etc. Excluded from the study were students whose course combinations could not be identified, including those who did not submit all assignments. As part of of the study the aim was to look into how different course combinations influenced academic performance, more specifically performance in English. Comparative analysis produced two distinct clusters. Students who combined English with either Music, Art or Computer Education performed better compared to those who combined English with SOSE (Study of Society and the Environment), Drama, Health and Physical Education, and Language and Linguistics. The results showed that there was little difference in the average marks obtained from different course combinations. The authors concluded that course combination had no effect on performance in the English course.

3 Research Approach

3.1 Instruments

There were two main tools used in the research: the Waikato Environment for Knowledge Analysis (WEKA) toolkit and the Success Or Failure Determiner (SOFD) tool, which was created specifically for this research. The predictive Apriori algorithm from WEKA was used to analyze the data set and this model was then integrated into the SOFD tool for further analysis and investigation.

3.2 Research Design

The research approach which was selected was the CRISP-DM methodology [3]. A modified version of this methodology is described in the phases presented below:

Data Understanding. The data used in this investigation was from the School of Computer Science at Wits. The data was extracted from the university database for the years 2006 to 2012. The data set consisted of a record for each student with the courses they enrolled into and their final academic outcome. The options for the the academic/progression outcome are shown in Table 1.

Data Processing. The data that was selected for the investigation consisted of all the courses enrolled into by that student and their final progression outcome. There were 564 student records. The data set required no cleansing. During the

Table 1. Overall Result Decision Codes

Decision Code	Meaning
PCD	Proceed
Q	Qualified
RET	return to year of study
CAN	canceled
MBP	Minimum requirements not met. Renewal of registration permitted by Examinations Committee (Proceed)
MBR	Minimum requirements not met. Renewal of registration permitted by Examinations Committee (Return)
MRNM	Minimum requirements not met. Needs permission to re-register
XXXX	Result depends on outcome of deferred examination(s)
****	Result not available
FTC	Failed to qualify

investigation the full data set was initially used then a reduced data set was used. The first was the full data set as extracted from the database. The second was a reduced data set where all the courses with less than 10 Computer Science students enrolled were removed from the full data set. Additionally all the second year courses were removed from the data set. This reduced the records from 564 to 528. Both these files were converted into .arff files which is the preferred WEKA format.

Modelling. The Predictive Apriori algorithm from the WEKA toolkit was applied on both the full data set and the reduced data set. The results from this are presented in Section 4. The models created from applying the Predictive Apriori algorithm to the data sets was then integrated into the SOFD toolkit. This is a GUI which has the WEKA General API embedded into it. This will allow for further investigation.

4 Results and Discussion

4.1 Full Data Set

Data Profiling. The initial data analysis was the profiling of the data set. This revealed the courses that were most preferred by Computer Science I students. These are shown in Figure 2. These courses exclude Mathematics I, which is a compulsory co-requisite of Computer Science I (COMS1000) and Computational and Applied Mathematics, which is a recommended co-requisite of COMS1000. Majority of the COMS1000 students were enrolled into the latter two courses. It is clear from Figure 2 that Physics and Economics are also very popular selections.

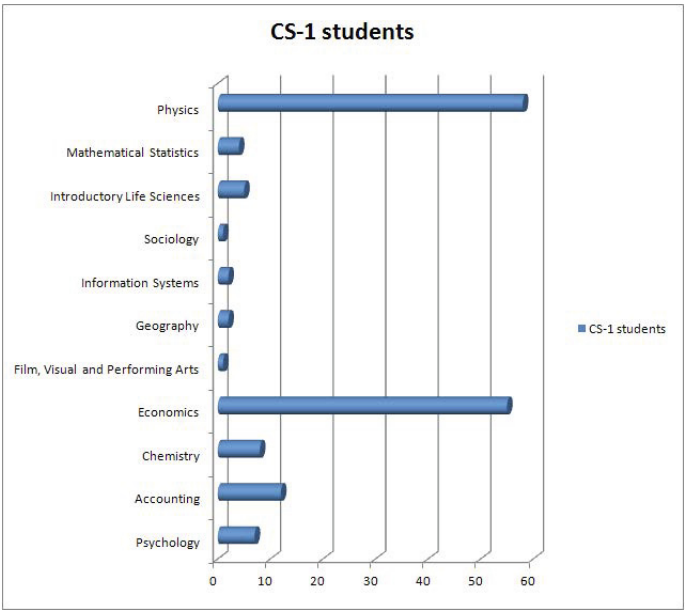

Fig. 2. Common Course Combinations

Analysis of Rules. Table 3 presents the rules produced after applying the Predictive Apriori algorithm on the full data set. All of the top ten rules have a good level of accuracy. From the ten rules, two progression outcomes are presented: 'RET'(which is failing the year) and 'PCD' (which is passing the year). Both these academic outcomes are described in Table 1. In this investigation the 'RET' would be seen as a negative rule and the 'PCD' was seen as the positive rule. 'RET' would result in a student returning to the same year of study. 'PCD' would result in the Computer Science student being allowed to proceed to the second year of study. From the 564 students in the data set: 240 of them obtained the 'PCD' outcome, 128 of them obtained the 'RET' outcome and the rest had the other outcomes mentioned in Table 1. Six of the top ten rules lead to an unsuccessful academic outcome for the Computer Science I students. The course codes presented in Table 3 are explained in Table 2.

Table 2. Course codes and titles

Course code	Course title
COMS1000	Computer Science
ELEN1002	Concepts of Design
INFO1003	Information Systems I
MATH1034	Algebra I
MATH1036	Calculus I
APPM1021	Applied Mathematics for Applied Computing I
PHYS1023	Physics for Applied Computing I

Table 3. Predictive Apriori Phase 1 Results

Attribute	Best Rules Generated	Accuracy
APPM1021	1. {APPM1021=yes, INFO1003=yes} $\Longrightarrow$ RET	0.83330
PSYC1002	2. {PHYS1023=yes, INFO1003=yes} $\Longrightarrow$ RET	0.83330
CHEM1013	3. {INFO1003=yes, ELEN1002=yes} $\Longrightarrow$ RET	0.83330
APPM1006	4. {BIOL1000=yes, CHEM1013=yes} $\Longrightarrow$ PCD	0.83330
ELEN1002	5. {BIOL1000=yes} $\Longrightarrow$ PCD	0.82353
ECON1000	6. {ECON100=yes, PSYC1002=yes} $\Longrightarrow$ PCD	0.79997
ECON2001	7. {ELEN1002=yes} $\Longrightarrow$ RET	0.75004
INFO1003	8. {APPM1006=yes, STAT1003=yes} $\Longrightarrow$ RET	0.75004
BIOL1000	9. {APPM1006=yes, ECON2001=yes} $\Longrightarrow$ PCD	0.74998
STAT1003	10. {APPM1006=yes, INFO1003=yes} $\Longrightarrow$ PCD	0.74998
PHYS1023		

Rule 4 and rule 5 both yield towards a positive academic outcome. Additionally, they both also have BIOL1000 as an antecedent in the rule. This is consistent with related literature which recognizes the link between Computer Science and Biology [2]. The area of Bio-informatics is an integration of Computer Science and Biology and one of the examples that show the association between the two courses. Another noteworthy finding in this result is that the INFO1003 course in three of the 'negative' rules. This is the Information Systems course offered by the Faculty of Commerce, Law and Management at Wits. It is clear from the results that COMS1000 students who select this course as one of their additional courses have a higher probability of failing their first year of study. This is an interesting result because Information Systems and Computer Science are usually branched off together. This result is therefore consistent with industry norms and related literature.

4.2 Reduced Data Set

The full data was then reduced by removing all the courses with less than ten Computer Science student enrollments and all courses which formed part of the second year of study. The new reduced data set is presented in Figure 3. From the reduced data set it is clear that most COMS1000 students enroll into Computational and Applied Mathematics, which is expected. Economics, Physics (Major) and Physics (Auxiliary) also enroll a lot of the COMS1000 students.

The top ten rules extracted from applying the Predictive Apriori algorithm to the reduced data set are shown in Table 4. Most of the resulting rules have good accuracy, however the last rule (tenth rule) has a low accuracy rate. It is also visible that most of the rules are 'positive', with only one rule being 'negative'. The first two rules with the highest accuracy are 'positive' rules with BIOL100/ Biological Sciences as a selection. This reaffirms the findings in the full data set which state that the selection of the Biological Science course by first year Computer Science students will increase the probability of them successfully completing their first year of study. The only other rule which had an

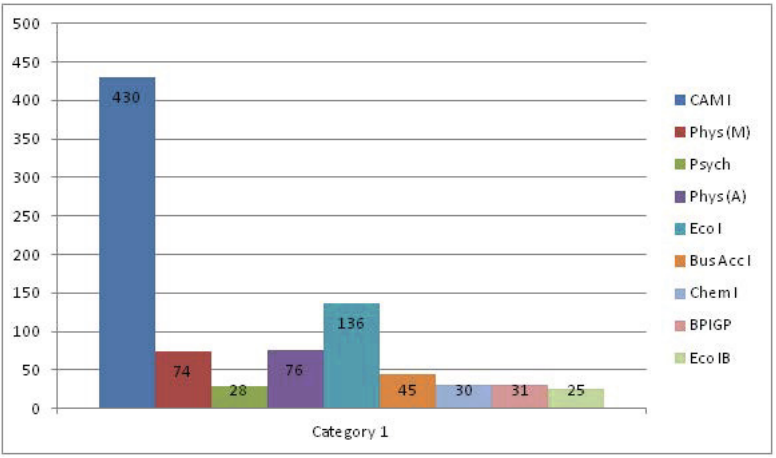

Fig. 3. Courses in Reduced Dataset

Table 4. Predictive Apriori Phase 2 Results

Attribute	Best Rules Generated	Accuracy
APPM1021	1. {BIOL1000=yes, CHEM1013=yes} $\Longrightarrow$ PCD	0.83333
PSYC1002	2. {BIOL1000=yes} $\Longrightarrow$ PCD	0.82353
CHEM1013	3. {ECON1000=yes, PSYC1002=yes} $\Longrightarrow$ PCD	0.80000
APPM1006	4. {ELEN1002=yes} $\Longrightarrow$ RET	0.76471
ELEN1002	5. {APPM1006=yes, STAT1003=yes} $\Longrightarrow$ PCD	0.75005
ECON1000	6. {APPM1021=yes} $\Longrightarrow$ PCD	0.73684
ECON2001	7. {PHYS1023=yes} $\Longrightarrow$ PCD	0.73684
INFO1003	8. {CHEM1013=yes} $\Longrightarrow$ PCD	0.69231
BIOL1000	9. {STAT1003=yes} $\Longrightarrow$ PCD	0.69231
STAT1003	10. {INFO1003=yes, APPM1021=yes,	0.25000
	PHYS1023=yes, ELEN1002=yes} $\Longrightarrow$ PCD	
PHYS1023		

accuracy of 0.8 and higher shows Psychology and Economics as resulting in a successful academic progression outcome. From Figure 3, it is shown that only a few Computer Science I students enroll into Psychology I. This should raise awareness to the course administrators and academics to advise more students to include Psychology I as one of their additional courses with Computer Science I. It is also significant that the only 'negative rule' was the selection of Concepts of Design as an additional course. This will require further investigation.

5 Conclusion and Future Work

During the registration process students select the courses they would like to enroll into. This decision is mainly made by the student but can be influenced

by advice from course coordinators and academics present during registration. Assistance from the academics could be based on research findings rather than previous experience and personal opinions. The aim of this investigation was to use data mining and historical data to find the association between course selection and academic performance. Results of applying the Apriori algorithm to historical data showed a relationship between course selection and academic performance. These results could be used to assist academics and course coordinators in preparation for registration.

Courses such as Biology, Economics and Psychology seem to have a better relationship with Computer Science I. Students who select these courses have a higher probability to be successful in their overall first year academic performance. These results are consistent with related literature.

This research showed two main things. Firstly, data mining can be used to find associations between courses and secondly, there are specific courses which when paired with Computer Science I yield successful academic performance. From this result, the SOFD tool can be considered as a recommendation tool for academics in the School of Computer Science. Course recommendation systems are becoming popular in the Educational Data Mining community. Most of the systems do focus on profiling the likes of students from historical data and using that to recommend courses to new students. This work uses academic performance as the key factor in assisting with the recommendation of courses to students. Future work can look at all the courses in Faculty of Science and considering which courses are best taken together to increase the first year pass rate. Although Computer Science I may be best matched with Economics, you may find another course which Economics I students enroll into, which is best matched with Economics. This can be taken further and investigated in other faculties and higher education institutions.

A limitation of this research is in the fact that the data set included students who are either repeating the first year of study or repeating some courses in the first year while enrolled in some second year courses. Although they are in the minority, they still influence the rules found. It is also noteworthy that the aim of this research was to find courses that would be good recommendations for first year students enrolling into Computer Science I, not necessarily to blame certain courses for failure in Computer Science I or first year.

Acknowledgments. I would like to acknowledge my supervisor Mr. Mike Mchunu for his assistance and guidance throughout the research. I would also like to thank Ms Bindu Cherian for assistance in the editing of this paper. And finally my husband, Landi Mashiloane, and family for their prayers and continuous support.

References

1. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering 32(1), 71–82 (2006)

2. Nilges, M., Linge, J.: Bioinformatics (2013), http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html
3. Shearer, C.: The crisp-dm model: The new blueprint for data mining. Journal of Data Warehousing 5(4), 13–22 (2000)
4. Sunita, B., Lobo, L.: Article: Best combination of machine learning algorithms for course recommendation system in e-learning. International Journal of Computer Applications 41(6), 1–10 (2012); Published by Foundation of Computer Science, New York, USA
5. Vannozzi, G., Della Croce, U., Starita, A., Benvenuti, F., Cappozzo, A.: Journal of neuroengineering and rehabilitation. Journal of Neuroengineering and Rehabilitation 1, 7 (2004)
6. Vialardi, C., Bravo, J., Shafti, L., Ortigosa, A.: Recommendation in higher education using data mining techniques. International Working Group on Educational Data Mining (2009)
7. Zailani, A., Tutut, H., Noraziah, A., Mustafa, M.D.: Mining significant association rules from educational data using critical relative support approach. Procedia - Social and Behavioral Sciences 28, 97 (2011), http://www.sciencedirect.com/science/article/pii/S1877042811024591; World Conference on Educational Technology Researches - 2011
8. Penn-Edwards, S.: They do better than us: A case study of course combinations and their impact on English assessment results. Educating: Weaving Research into Practice (2004)