Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.3, No.8, 2013

Analyzing Undergraduate Students' Performance in Various Perspectives using Data Mining Approach

Anwar M. A.[1*] Naseer Ahmed[2]

1. College of Engineering and Computing, Al Ghurair University, Dubai Academic City, PO Box 37374, United Arab Emirates
2. Institutional Research and Effectiveness, Al Ghurair University, Dubai Academic City, PO Box 37374, Dubai, United Arab Emirates

* E-mail of the corresponding author: anwar@agu.ac.ae

Abstract
The data mining provides better insight rather than the predefined queries or reports for quality enhancement and improvement of an academic program to extract hidden knowledge in students' performance in various courses. This paper presents data mining approach applied to discover students' performance patterns in two different perspectives (a) supervised and unsupervised assessment instruments and (b) discover students' performance patterns in mathematics, English, and programming courses in an engineering degree program. The interesting patterns emerging from both analytic studies offer helpful and constructive suggestions for the improvement and revision of assessment methodologies, restructuring the curriculum, and modifying the prerequisites requirements of various courses.
Keywords: Association Rules, Supervised and Unsupervised Assessment, Educational Data Mining

1. Introduction

The past several decades have witnessed a rapid growth in the use of data and knowledge mining as a means by which academic institutions extract useful hidden information in the student result repositories in order to improve students' learning processes. Data mining (also called data or knowledge discovery) is the method of analyzing data from different perspectives to discover interesting and helpful information. The information gained through data mining has been effectively used in various sectors ranging from finance, agriculture to health and education. There are many data mining tools (Weka 2012), (XLMiner 2013), (KNIME 2013) available that allow users to analyze data from many different aspects, categorize it, and discover the identified relationships. Technically, data mining is a technique of finding correlations or patterns among many fields in large databases. Educational data mining is fast becoming an interesting research area which allows researcher to extract useful, previously unknown patterns from the educational databases for better understanding, improved educational performance and assessment of the student learning process (A.Y.K. Chan et al. 2007) It facilities the exploration of unique information from students' result database in academic institutions.

The essential part of curriculum of a computer science and engineering degree program is English, mathematics, and programming. The programming is taught at introductory level, intermediate level and advanced levels. The programming often requires expertise in many different subjects, including knowledge of the application domain, analytical skills, and comprehension of the program requirement specification. One of the main objectives of the calculus courses is to develop analytical skills in the student whereas the English courses develop the comprehension of the problem statements in programming or any other area. The student's performance in a course is assessed through a variety of assessment instruments i.e. assignments, projects, laboratory work, semester end examinations etc. Some of these assessment instruments are unsupervised such as assignments, homework, and projects for which students are at a liberty to take help from textbooks or reference material etc. The assessments in this category are an essential part of learning process and can be regarded as a mean of preparing students for supervised assessments, for example, tests, presentations, oral examinations. The unsupervised assessments are administered under the constant vigilance of a teacher or an examiner with no outside help or assistance. It is generally expected that students performing well in an unsupervised assessment would also perform well in a supervised assessment. Similarly, it is also a perception that if a student's performance is better in a prerequisite course of a course under assessment or in mathematics and English then he will also perform better in the a programming course.

From the stand view of the e-learning scholars, data mining techniques have been employed to solve different problems in the educational environment. The selection of data mining tools and techniques mostly depends on the scope of the problem and the expected results from the analysis. For example, a classification approach is used (B. Minaei Bidgoli 2003) to classify students to predict their final year performance based on different parameters derived from the data in an educational web-based system. A clustering algorithm is used (G.J. Tsai) to categorize students with similar behavioral characteristics. Association rule mining techniques have frequently been used to solve educational problems and carry out critical analysis in an academic environment

for improving the learning process of student. These efforts are carried out in order to raise the standards and administration of educational processes by investigating the learning systems, learning resources arrangements, and students' results, curriculum restructuring, and institutional websites (R. Damasevicius et al. 2009), (Talavera, L. et al. 2004), (S. Z. Erdogan et al. 2005). In one study (Anna Katerina et al. 2010) clustering and association rule mining techniques have been applied to students' data to mine the common factors affecting the learners' performance that can be utilized as a base for providing some clues and hints for future learners. In another study (David H. et al. 2010), students' actions logged during tutor session have been categorized, binned, and symbolized to discover student behavior patterns.

In this paper we used *Appriori* algorithm to mine rules in (a) supervised and unsupervised assessment results and (b) programming, mathematics, and English results. The rules meeting the predefined support and confidence are mined to expose the hidden knowledge from the available student assessment data. These mined rules are analyzed to review the existing assessment processes and recommend constructive actions to academic planners. In section 2, we present relevant information about knowledge discovery process along with the data mining and association rule that we have used for the discovery of hidden knowledge. The results of the analysis and the rules discovered from the present study are discussed in section 3. The conclusions of our work are given in section 4.

2. Proposed system

Knowledge Discovery (KD) process is one of the basic concepts of the field of Knowledge Discovery and Data mining (KDD). Figure 1 illustrates the knowledge discovery employed in the present study that we have adapted from (Anna Katerina Domminguez et al. 2010). Solid-line arrows represent various important data processing steps leading towards the knowledge discovery whereas dotted-line arrows show the steps that may form an iterative cycle in order to refine the knowledge discovery process.

2.1 Selecting Mining Frequent Patterns and Associations

The association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes' value conditions that occur frequently together in a given dataset (Jiawei Han et al 2006). The preliminaries necessary to understand for performing data mining on any data are discussed below.

Let $I = \{I_1, I_2, I_3, ..., I_m\}$ be a set items. Let D, the task relevant data, be a set of database transactions where each transaction $T \subseteq I$. Each transaction is an association with an identifier, called transaction identification (TID). Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I, B \subset I$, and $A \cap B = \phi$.

Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A that also contains B.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. If the relative support of an itemset I satisfies a prescribed minimum support threshold, then I is a frequent itemset.

The association rule mining can be viewed as a two-step process:

a. Find all frequent itemsets: Each of these itemsets will occur at least as frequently as a predetermined minimum support count.
b. Generate strong association rules from the frequent itemsets: The rules must satisfy minimum support and confidence. These rules are called strong rules.

2.2 Apriori Algorithm

Apriori is a seminal algorithm proposed by (R. Agarwal et al. 1994) for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The following lines state the steps in generating frequent itemset in Apriori algorithm.

Let *Ck* be a candidate itemset of size *k* and *Lk* as a frequent itemset of size *k*. The main steps of iteration are:

Find frequent set L_{k-1}

Join step: C_k is generated by joining L_{k-1} with itself (cartesian product L_{k-1} x L_{k-1})

Prune step (apriori property): Any $(k - 1)$ size itemset that is not frequent cannot be a subset of a frequent k size itemset, hence should be removed

Frequent set L_k has been achieved

2.3 Task Relevant Data Collection

We analyzed the available result data of three courses in two perspectives described earlier. The score of each student was transformed into transactions (assignment, laboratory work, class test, final examination) for supervised and unsupervised perspective and (TID, programming grade, mathematics grade, and English grade) for analyzing correlation of results in three courses. The student ID will serve as TID, however, it is not included while applying data mining algorithm.

2.4 Data Preprocessing

The real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results (Han, J. et al. 2005). Therefore, data preprocessing is an important task in data mining. The data we used was in the percentages as discussed above and needed to be transformed to same level of achievement in each assessment. Hence, all scores in different assessment instruments were transformed to a number calculated out of 100. Table 1 show the data used in the analysis after preprocessing. Symbols A, Q, FL, T, and FE are used to identify assignments, quizzes, laboratory work, test, and final examination respectively. The result data in each assessment instrument was preprocessed to grades (Stage-1) A(>= 90), B(>= 80), C(>=70), D(>= 60), and F(<60) as shown in Table 1. These grades were concatenated with the type of assessment for example an A-A represents A grade in assignment and FE-B represents a B grade in final examination. The final pre-processed form (Stage-2) of assessment transaction, for example, is highlighted in the left part of Table 1 by a rectangular box.

2.5 Data Cleaning

It is fundamental truth that incorrect or inconsistent data can lead to false conclusions and hence wrong inferences and decisions. Therefore, high quality data needs to pass a set of quality criteria; accuracy, integrity, completeness, validity, consistency, uniformity, density, and uniqueness. Data cleaning routines attempts to fill in missing values, smooth out noise, and correct inconsistencies in the data. There are a number of data cleaning techniques (Han, J. et al. 2005) in the literature such as fill missing values, binning, regression, and clustering. We used the following criteria to clean our data:

a. If a student did not sit in the final examination then zero is entered in his score. We removed all such tuples from our result data.
b. If a student is absent in one or two assessment instrument then his score was replaced by average of the students score in that assessment.
c. If a student is absent in more than two assessment instruments then all such tuples were removed.

Table 1: Transformed assessment data

For perspective (a)

(a) Preprocessed data (Stage 1)					(b) Preprocessed data (Stage 3)				
A	**Q**	**FL**	**T**	**FE**	**A**	**Q**	**FL**	**T**	**FE**
A	D	F	C	D	A-A	Q-D	FL-F	T-C	FE-D
F	F	D	C	D	A-F	Q-F	FL-D	T-C	FE-D
B	F	F	F	F	A-B	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
A	F	D	C	F	A-A	Q-F	FL-D	T-C	FE-F
B	B	D	C	B	A-B	Q-B	FL-D	T-C	FE-B
A	C	F	C	C	A-A	Q-C	FL-F	T-C	FE-C
A	C	F	F	F	A-A	Q-C	FL-F	T-F	FE-F
B	F	F	F	F	A-B	Q-F	FL-F	T-F	FE-F
A	D	F	C	F	A-A	Q-D	FL-F	T-C	FE-F
B	D	F	D	F	A-B	Q-D	FL-F	T-D	FE-F
B	F	F	D	F	A-B	Q-F	FL-F	T-D	FE-F
B	C	F	C	F	A-B	Q-C	FL-F	T-C	FE-F
A	F	F	F	F	A-A	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
B	F	F	F	F	A-B	Q-F	FL-F	T-F	FE-F

For perspective (b)

Preprocessed data (Stage 2)				Transformed data (Stage 3)		
P	M	E-1	E-2	P	M	E
C	C	C	A	P-C	M-C	E-C
C	C	B	A	P-C	M-C	E-A
C	C	B	B	P-C	M-C	E-B
C	C	C	B	P-C	M-C	E-A
A	A	A	A	P-A	M-A	E-A
C	C	C	B	P-C	M-C	E-B
D	D	C	B	P-D	M-D	E-C
D	C	C	C	P-D	M-C	E-C
C	C	C	A	P-C	M-C	E-C
C	D	C	D	P-C	M-D	E-D
F	D	C	F	P-F	M-D	E-D
C	C	A	B	P-C	M-C	E-B
B	B	A	A	P-B	M-B	E-A
C	C	B	B	P-C	M-C	E-A
B	B	A	B	P-B	M-B	E-B

The final pre-processed form of assessment data using to mine association for perspective (b) is also shown in

Table 1 where P represents grades in programing course, M represents grades in calculus course, E is used for English comprehension course and technical writing courses. The abbreviation of courses is used as prefix to the grades in these courses i.e. P-C represents C grade in programming, as highlighted by a rectangular box.

3. Results and Rules Analysis

The association rules mined from supervised and unsupervised students' results are shown in Table 2 whereas the association rules from the results of programming, mathematics, and English courses are shown in Table 3. A number of uninteresting rules have been excluded from Tables 2 and 3 due to the limitation of the space. The association rules depicted in Tables 2 and 3 are mined using a data mining tool (XLMiner 2013). This tool allows mining the association rules by setting various minimum support thresholds. It is observed that by lowering the minimum support threshold there is a marked increase in the number of association rules generated by XLNimer tool in both perspectives.

Table 2: Association rules mined; minimum support 7 and confidence 75%

Rule #	Antecedent	Consequent	Support	Conf. %
1	A-A	FL-F	8	87.5
2	A-A	FE-F	8	75
3	A-B, FE-F	FL-F	6	100
4	A-B, FL-F	FE-F	6	100
5	A-B	FE-F, FL-F	7	85.71
6	A-B	FE-F	7	85.71
7	A-B	FL-F	7	85.71
8	A-F	Q-F	6	100
9	FE-F, FL-F, Q-F	T-F	13	84.62
10	FL-F, Q-F	FE-F, T-F	13	84.62
11	FL-F	FE-F	21	80.95
12	Q-F	FE-F	15	93.33
13	Q-F	FL-F	15	86.67
14	T-F	FE-F, FL-F, Q-F	12	91.67

Table 3: Association rules mined; minimum support 8 and confidence 86%

Rule #	Antecedent	Consequent	Support	Conf. %
1	P-A	M-A	9	100
2	E-A & P-A	M-A	9	100
3	P-A	M-A & E-A	9	100
4	P-B	M-B	12	100
5	P-C	M-C	38	88
6	E-A & P-C	M-C	10	100
7	E-B & P-C	M-C	17	100
8	E-C & P-C	M-C	10	86
9	M-A	E-A	9	100
10	P-A	E-A	9	100
11	M-A & P-A	E-A	9	100
12	M-A	E-A & P-A	9	100
13	M-A	P-A	9	100
14	M-A & E-A	P-A	9	100
15	M-B	P-B	12	100
16	M-C & E-A	P-C	10	86

The analysis of the generated rules perspective (a) presented in Table 2 shows that rule 1 (support = 8, confidence 87.5%) indicates that students who performed excellent in assignment failed to perform even satisfactorily in the final laboratory examination. The rule 2 (support = 8, confidence 75%) is also strong and extracts knowledge that students who performed excellent in assignment failed in final examination. Similarly, rule 6 (support = 7, confidence 85.71%) and rule 7 (support = 7, confidence 85.71%) show that students who performed very good i.e. scored B grade failed to score similar grades in the final examination and the final laboratory work, respectively. A similar trend is observed in the rules generated with minimum support 3 and minimum confidence greater than 80%. We could not find a single rule with minimum support 3 and 6 and minimum confidence greater that 75% that verifies that students performing excellent in the unsupervised assessment instruments surely performed well in the supervised assessment instruments. The discovered rules are strangely contradictory to the fact that if a student's performance is excellent in the unsupervised assessment (homework or assignments) then he/she must perform better in the supervised assessment instruments such as tests, laboratory works, and/or final examination. This could be due to a variety of reasons; (i) assignments or homework were not developed properly, (ii) there might be an impedance mismatch in the unsupervised and supervised assessments, (iii) the students were not able to apply the knowledge and skills gained through unsupervised assessments in the final laboratory examination and/or class tests or final examination, (iv) the students might have copied the assignments and homework either from the resources available on the Internet or from their friends. This might be another possible explanation for the strange results uncovered from this study.
The analysis of the generated rules perspective (b) presented in Table 3 show that if a student's performance is excellent in mathematics or mathematics and English then he/she must perform better in the programming courses but excellent performance in English alone does not guarantee same performance in programming course. This could be due to the reason that the students understand the problem by translating the problem statement in English to their native language. The rules discovered in this perspective do confirm many findings from previous studies using non KDD approaches (Talavera, L. et al. 2004). There is a positive correlation between the students' problem solving ability and their programming performance.

4. Conclusion

The paper presented the potential use of one of the data mining approaches called association rule mining algorithm in enhancing the quality and experience of students' performances in higher education. The analysis reveals that there are more students who got excellent grades in supervised assessment but failed to attain similar level of performance in the unsupervised assessments and if a student's performance is excellent in mathematics or mathematics and English then he/she must perform better in the programming courses but excellent performance in English alone does not guarantee same performance in programming course. All these and alike hidden patterns could serve as an important feedback for instructors, curriculum planners, academic managers, and other stakeholders in making informed decisions for evaluating and restructuring curricula and associated assessment methodologies with a view to improve students' performance in supervised and unsupervised assessment instruments.

5. Acknowledgements

The authors wish to acknowledge the financial support provided by the Al Ghurair University.

References

A.Y.K. Chan, K.O. Chow, and K.S. Cheung (2007). "Online Course Refinement through Association Rule Mining", Journal of Educational Technology Systems, Volume 36, Number 4/2007 – 2008, pp 433 – 444.
Anna Katerina Domminguez, Kalina Yasef, and James R. Curran (2010). "Data Mining for Individualized Hints in eLearning", in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA.
B. Dogan and A. Y. Camurcu (2007). "Association Rule Mining form an Intelligent Tutor", Journal of Educational Technology Systems, Volume 36, Number 4/2007 – 2008, pp 444 – 447.
B. Minaei Bidgoli, B.A. Kashy, G. Kortemeyer, and W. F. Punch (2003). "Predicting Students Performance: an application of data mining methods with the educational web-based system LON-CAPA", in proceedings of ASEE/IEEE Frontier in Education Conference, Boulder, CO: IEEE.
David H. Shanabrook, David G. Cooper, Beverly Park Woolf, and Ivan Arroyo (2010). "Identifying High-Level Student Behavior Using Sequence-based Motif Discovery", in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA.
G.J. Tsai, S.S. Tseng, and C.Y. Lin. "A Two Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment". In proceedings of the Alexandrov, V.N. et al. (eds.)
Han, J., Kamber, M., Pei, J. (2005). "Data Mining: Concepts and Techniques" The Morgan Kaufmann Series in Data Management Systems, (ISBN: 9781558609013) 2nd Edition.

Jiawei Han and Micheline Kamber (2006). Data mining: concepts and techniques, Morgan Kaufmann.
KNIME (2013). http://www.knime.org/ (May 2013)
R. Damasevicius (2009). "Analysis of Academic Results for Informatics Course Improvement using Association Rule Mining". Information Systems Development towards a Service Provision Society. ISBN 978-0-387-84810-5 (print), pp 357 – 363, published by Springer USA.
S. Z. Erdogan, m. Timor. (2005) "A Data Mining Application in a Student Database". Journal of Aeronautics and Space Technologies, Vol. 2, Number 2., pp 53 – 57.
Talavera, L., Gaudioso, E. (2004) "Mining Students Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces". In proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI , Valencia, Spain.
R. Agrawal and R. Srikant (1994). Fast algorithms for mining association rules in large databases. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California.
Weka (2013). http://www.cs.waikato.ac.nz/ml/weka/ (May 2013)
XLMiner (2013). (http://www.resample.com/xlminer/index.shtml (May 2013)

Anwar M. Abaidullah is working as Assistant Professor and Deputy Dean of College of Engineering and Computing in Al Ghurair University, UAE. He received his Doctorate of Engineering with specialization in object-oriented databases from Kyushu Institute of Technology, JAPAN in 2001. Since 2001, he has been affiliated with renowned universities in GCC and Pakistan.

Naseer Ahmed is working as Director of Institutional Research and Effectiveness at Al Ghurair University, UAE. He received his PhD in Physics from Heriot Watt University, UK.