Knowledge Discovery from Students' Result Repository: Association Rule Mining Approach

Oladipupo O.O. frajooje@yahoo.com
College of Science and Technology, Department of Computer and Information Sciences
Covenant University
Ota,PMB 1023, Nigeria

Oyelade O.J. ayo2006ola@yahoo.com
College of Science and Technology, Department of Computer and Information Sciences
Covenant University
Ota,PMB 1023, Nigeria

Abstract

Over the years, several statistical tools have been used to analyze students' performance from different points of view. This paper presents data mining in education environment that identifies students' failure patterns using association rule mining technique. The identified patterns are analysed to offer a helpful and constructive recommendations to the academic planners in higher institutions of learning to enhance their decision making process. This will also aid in the curriculum structure and modification in order to improve students' academic performance and trim down failure rate. The software for mining student failed courses was developed and the analytical process was described.

Keyword: Association rule mining, Academic performance, Educational data mining, Curriculum, Students' Result Repository.

1. INTRODUCTION

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process [1]. It is concerned with developing methods for exploring the unique types of data that come from educational environment which include students' results repository.

Students' result repository is a large data bank which shows the students raw scores and grades in different courses they enrolled for during their years of attendance in the institution. Student performance score is basically determined by the sum total of the continuous assessment and the examination scores. In most institutions the continuous assessment which includes various assignments, class tests, group presentations is summed up to weigh 30% of the total score while the main semester examination is 70%. To differentiate different students' performances and scores a set of alphabetic grade is identified to represent the score ranges such as 70-100

as “A”, 60-69 as “B”, 50 to 59 as “C” and 45-49 as “ D” and < 45 as “F”. Any score < 45 is regarded as a fail performance. This grade representation is different from one higher institution to another.

From the standpoint of the e-learning scholars, data mining techniques is said to have been applied to solve different problems in educational environment which includes Students’ classification based on their learning performance; detection of irregular learning behaviors; e-learning system navigation and interaction optimization; clustering according to similar e-learning system usage; and systems’ adaptability to students’ requirements and capacities and so on. [2] The choice of data mining tool is mostly determined by the scope of the problem and the expected analysis result.

In [3] an approach to classify students in order to predict their final year grade based on the features extracted from logged data in an educational web-based system was reported. Data mining classification process was used in conjunction with genetic algorithm to improve the prediction accuracy. Also, in [4] student data was mined to characterize similar behavior groups in unstructured collaboration using clustering algorithms. The relationship between students’ university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques in [5]. Fuzzy logic concept was not behind in the field of educational data mining [6,7,8,9], for instance a two-phase fuzzy mining and leaning algorithm was described in [10], this is an hybrid system of association rule mining apriori algorithm with fuzzy set theory and inductive learning algorithm to find embedded information that could be fed back to teachers for refining or reorganizing the teaching materials and test. Association rule mining technique has also been used in several occasions in solving educational problems and to perform crucial analysis in the educational environment. This is to enhance educational standards and management such as investigating the areas of learning recommendation systems, learning material organization, student assessments, course adaptation to the students’ behaviour and evaluation of educational web sites [1,11,12 13,14]. In [12] a Test Result Feedback (TRF) model that analyses the relationships between students’ learning time and the corresponding test results was introduced. Knowledge Discovery through Data Visualization of Drive Test Data was carried out in [15]. Genetic algorithm as Ai technique was for data quality mining in [16] Association rule mining was used to mining spartial Gene Expressing [17] and to discover patterns from student online course usage in [14] and it is reported that the discovered patterns from student online course usage can be used for the refinement of online course. Robertas, in [18] analysed student academic results for informatics course improvement, rank course topics following their importance for final course marks based on the strength of the association rules and proposed which specific course topic should be improved to achieve higher student learning effectiveness and progress.

In view of the literature, it is observed that different analysis has been done on students’ result repository but the failed courses in isolation has never been analysed for hidden and important patterns, which could be of a great importance to academic planners in enhancing their decision making process and improving student performance. In order to bridge this gap, this paper presents an analysis of students’ academic failed courses in isolation using association rule mining. This is to discover the hidden relationships that exist between different students failed courses in form of rules. The generated rules are analysed to make useful and constructive recommendations to the academic planners. This promised to enhance academic planner’s sense of decision making and aid in the curriculum structure and modification which in turn improve students’ performance and trim down failure rate.

1.1 Association Rule mining

Association rule mining associates one or more attributes of a dataset with another attributes, to discover hidden and important relationship between the attributes, producing an if-then statement concerning attribute values in form of rules. (19,20). The formal definition of association rule

mining is : Let I = $\{i_1,i_2,\ldots,i_m\}$ be a set of literals called items and D be a set of transactions where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$. Association rule mining process could be decomposed into two main phases to enhance the implementation of the algorithm. The phases are:

1. Frequent Item Generation: This is to find all the itemsets that satisfy the minimun support threshold. The itemsets are called frequent itemsets.
2. Rule Generation: This is to extract all the high confidence rules from the frequent itemsets found in the first step. These rules are called strong rules.

Over the years different algorithms have been proposed in the literature that implement the two phases of association rule mining [21]. In this paper the traditional Apriori algorithm is implemented to generate the hidden patterns from the students' failed courses dataset which when analysed will serve as a strong convincing recommendation to academic planning department in institutions of learning for curriculum structure and modification in other to improve the students' performances and minimize failure rate percentage.

1.2 Definition of terms

Association rules: An association rule is an implication expression of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and X and Y are disjoint itemsets, i.e $X \cap Y = \phi$. The strength of an association rule can be measured in terms of its support and confidence. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence *c* and support *s*, if c% of the transactions in D that contains X also contains Y, and s% of transactions in D contains $X \cup Y$. Both the antecedent and the consequent of the rule could have more than one Item. The formal definitions of these two metrics are:

$$Support, s\ (X \Rightarrow Y) = \frac{\Sigma(X \cup Y)}{N} \quad (1)$$

$$Confidence, c\ (X \Rightarrow Y) = \frac{\Sigma(X \cup Y)}{\Sigma X} \quad (2)$$

Example 1: Consider a rule { CSC111,CSC121} $\Rightarrow$ {CSC211}. If the support count for {CSC111, CSC121, CSC211} is 2 and the total number of transactions is 5 then, the rule's support is 2/5 = 0.4. The rule's confidence is obtained by dividing the support count for {CSC 111, CSC121, CSC211} by the support count for {CSC111, CSC121}. If there are 3 transactions that contain CSC111, CSC121 then, the confidence for this rule is 2/3 = 0.67. If the minimum rule support is 0.3 and minimum confidence is 0.5, then, the rule
{ CSC111,CSC121} $\Rightarrow$ {CSC211} is said to be strong, that is; the interestingness of the rule is high.

1.3 Justification for Support and Confidence measure

Support is an important measure because a rule that has a very low support may occur by chance. A low support rule in this context is likely to be uninteresting from the academic perspective because such a failure combination might come accidentally and it might not be profitable to enhance academic planner decision. Also, confidence on the other hand, measures the reliability of the inference made by a rule. So, the higher the confidence, the more frequent the failed courses appear together within the database.

2. METHODOLOGY:

2.1 Development of an Apriori Algorithm

The algorithm starts by collecting all the frequent 1-itemsets in the first pass based on the minimum support. It uses this set (called L_1) to generate the candidate sets to be frequent in the next pass (called C_2) by joininig L_1 with itself. Any item that is in C_1 and not in L_1 is eliminated from C_2. This is achieved by calling a function called 'apriori-gen'. This reduces the item size drastically. The algorithm continues in the same way to generate the C_k, of size k from the large itemsets of k-1, then reduces the candidate set by eliminating all those items in k-1 with support count less than minimum support. The algorithm terminates when there are no candidates to be counted in the next pass. Figure 1 shows the general Pseudocode for association rule mining and Figure 2 shows the traditional apriori algorithm, while figure 3 , shows the algorithm for 'apriori-gen' function called for candidate generation and elimination of non-frequent itemset.

Step 1:	Accept the minimum support as minsup and minimum confidence as minconf and the student failed course as the input data set.
Step 2:	Determine the support count for all the item as s (courses under consideration).
Step 3:	Select the frequent items; item with s ≥ minsup
Step 4:	The set candidate k- item is generated by 1- extension of the large (k-1) itemsets generated in step3
Step 5:	Support for the candidate k-itemsets are generated by a pass over the database.
Step 6;	Itemset that do not have minsup are discarded and the remaining itemsets are called large k-itemsets.
Step 7 :	The process is repeated until no more large item.
Step 8:	The interesting rules are determined based on the minimum confidence.

FIGURE 1: General Pseudocode for Association Rule Mining

2.2 Apriori Candidate Generation

The apriori-gen is a function called in line 3 of the algorithm1. It takes as agument L_{k-1} , the set of all large (k-1) itemsets. It returns a superset of the set of all large k-itemsets. The description of the function is given in algorithm 2.

2.3 Rule generation

After all the frequent itemsets have been generated then the rules are determined. In rule generation, we do not have to make additional passes over the data set to compute the support of the candidate rules. All needed is to determine the confidence of each rule by using the support counts computed during frequent itemsets generation.

Find frequent set L_{k-1}.

Join Step.

C_k is generated by joining $L_{k-1}$with itself

Prune Step.

Any $(k-1)$ – itemset that is not frequent cannot be a subset of a frequent k - itemset, hence should be removed.

where

(C_k: Candidate itemset of size k)

(L_k: frequent itemset of size k)

FIGURE 2: Frequent itemset generation of the Traditional Apriori Algorithm [21]

Apriori (T, ε)

$L_1 \leftarrow \{$*large 1-itemsets that appear in more than* ε *transactions* $\}$

$k \leftarrow 2$

while $L_{k-1} \neq \varnothing$

$C_k \leftarrow$*Generate*(L_{k-1})

for transactions $t \in T$

$C_t \leftarrow$Subset(C_k,t)

for candidates $c \in C_t$

$\text{count}[c] \leftarrow \text{count}[c] + 1$

$L_k \leftarrow \{c \in C_k | \ \text{count}[c] \geq \varepsilon\}$

$k \leftarrow k + 1$

return $\cup_k L_k$

FIGURE 3: Function for generating candidate itemset [21]

3. Result and Rule analysis

In most literature authors focus on the students' aggregate performances; Grade Point Average [12,13,14, 18] and their findings are useful majorly for prediction, which might not really improve the low capacity students' performances. In this research the association rule mining analysis was performed based on students' failed courses. This identifies hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. Figure 4 shows a snapshot for association mining process interface. In this work it is observed that the lower the items minimum support, the larger the candidate generated. This adversely affects the complexity of the system. For instance, in figure 4, if the item minimum

support is 3 and the rule confidence is 0.5, we have 19 frequent itemsets and 114 rules are generated. Table 2 show the relationship between the minimum support, minimum confidence and the generated rule and figure 5 gives the graphical representation.

It was observed that the execution time is also inversely proportional to minimum support, since it increases as minimum support decreases, which confirmed increase in system complexity and response time as the minimum support decreases as shown on table 2. With all these observations it shows that to have a less complex system and a constructive, interesting and relevant patterns the minimum confidence and support should be large enough to trash out coincidence patterns. Table 3 also displayed some of the rules generated.

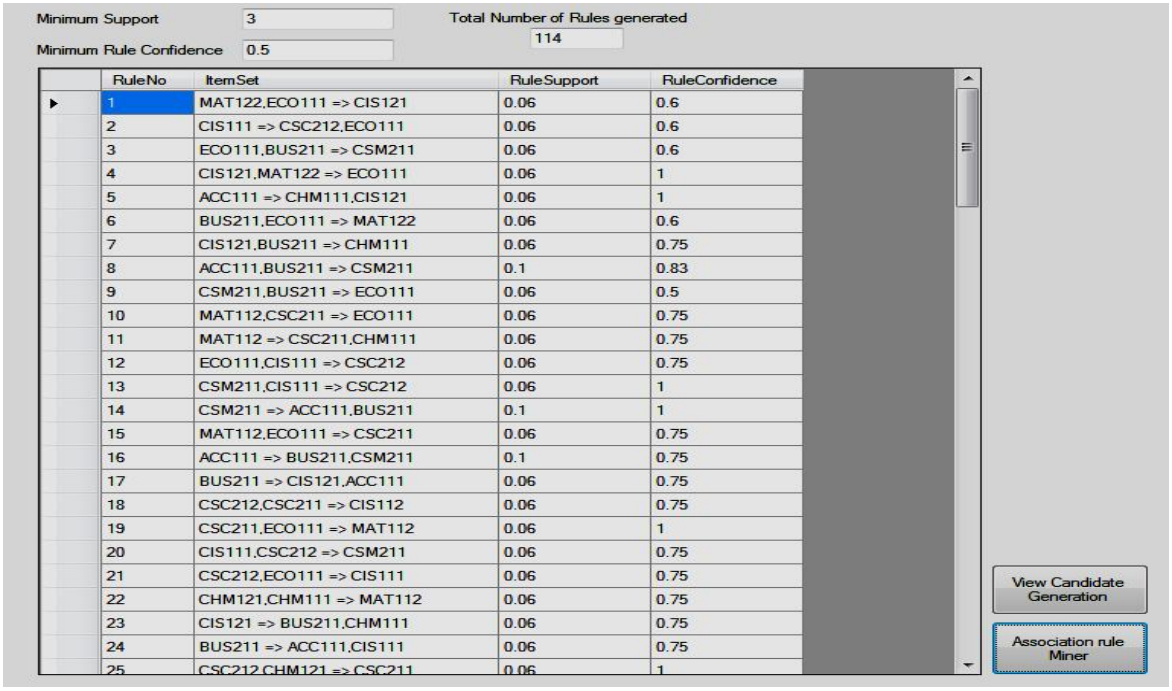

FIGURE 4 :A snapshot for Association rule mining process interface

TABLE 1: Relationship between minimum confidence , minimum support and number of generated rules.

	Minimum Item(s) support = 3 Average Exe.Time = 6sec		***Minimum item(s) support = 3 Average Exe.Time = 14sec***		***Minimum item(s) support = 3 Average Exe.Time = 40sec***	
Min.Conf.	**#Rules**	**#of freq. Itemset**	**#Rules**	**#of freq. Itemset**	**#Rules**	**#of freq. Itemset**
50%	6	1	114	19	855	152
60%	6	1	97	19	631	152
70%	6	1	82	19	345	152
80%	6	1	51	19	306	152
90%	2	1	49	19	301	152
100%	2	1	49	19	301	152

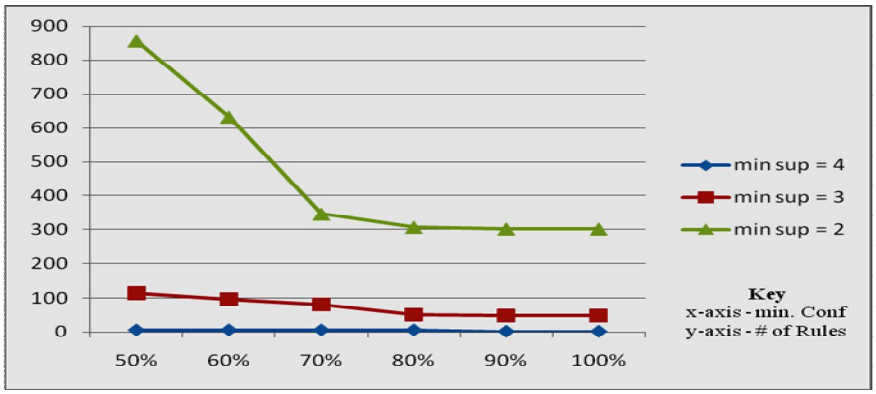

FIGURE 3: Graphical representation of effect of minsup, minconf on #of rules

3.1 Rule Analysis

Table 2 shows an instance of the rule generated from the simulation. All the rules with confidence 1, are very strong rules, which implies that if a student failed the determinant (antecedent) course(s), such student will surely fail the dependent (consequent) course(s). Such rules should not be overlooked in curriculum structure. Also if the rule support is higher, it means that all the courses involved are failed together by most of the considered students. From rule number 1 one can deduce that MAT 122, ECO 111⇒CIS 121 with (s =0.06, c = 0.6). This indicates that, the probability that every student that fails MAT 122, ECO 111 will also fail CIS 121 is 0.6. This type of rule is not very strong; in some cases it might be overlooked but notwithstanding, the academic planner can still take it into consideration. In that case, MAT 122 and CIS 121 should not be taken in the same semester. This kind of failure can be minimized if one becomes a prerequisite to another. That is, if a student has not passed MAT 122 they will not be allowed to register for CIS 121. Also, we have from rule 8, a strong rule such that ACC111, BUS211 => CSM211 with (s = 0.1, c= 0.83). ACC 111 is introduction to accounting; BUS 211 is introduction to Business and CSM 211 Mathematical method 1. The first two courses are compulsory courses for the Management Information System students. ACC111 is a 100 level first semester course while the other two are 200 level first semester courses. This implies that a student that fails ACC 111 in 100 level should not be allowed to register for BUS 211 or CSM 211 and if possible, the two, so as to avoid multiple failure.

With all these observations, if academic planners can make use of the extracted hidden patterns from students' failed causes using association rule mining approach, it will surely help in curriculum re-structuring and also, help in monitoring the students' ability. This will enable the academic advisers to guide students properly on courses they should enroll for. This, eventually, tends to increase the student pass rate.

TABLE 2: An instance of rule generated with support and confidence

RuleNO	Rule	Rule Support	Confidence
1	MAT122,ECO111 => CIS121	0.06	0.6
2	CIS111 => CSC212,ECO111	0.06	0.6
4	CIS121,MAT122 => ECO111	0.06	1
5	ACC111 => CHM111,CIS121	0.06	1
6	BUS211,ECO111 => MAT122	0.06	0.6
7	CIS121,BUS211 => CHM111	0.06	0.75
8	ACC111,BUS211 => CSM211	0.1	0.83

4. Conclusion, Recommendation and Future Work

This study has bridge the gap in educational data analysis and shows the potential of the association rule mining algorithm for enhancing the effectiveness of academic planners and level advisers in higher institutions of leaning. The analysis was done using undergraduate students' result in the department of Computer Science from a university in Nigeria. The department offers two programmes; Computer Science and Management Information Science. A total number of 30 courses for 100 level and 200 level students are considered as a case study. The analysis reveals that there is more to students' failure than the students' ability. It also reveals some hidden patterns of students' failed courses which could serve as bedrock for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate. To adopt this approach a larger number of students should be considered from the first year to the final year in the institution. This will surely reveal more interesting patterns. Also, the min. confidence should be of a higher percentage to be able to have more relevant and constructive rules. In future applications, in order to improve the comprehensibility and applicability of the association rules, it will be very useful to also provide an ontology that would describe the content of the courses which will allow the academic planners to understand better the rules that contain concepts related to the analysed domain.

References

1. B. Dogan, A. Y. Camurcu. "Association Rule Mining from an Intelligent Tutor" Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 447, 2008
2. F. Castro, A. Vellido, A. Nebot, and F. Mugica. "Applying Data Mining Techniques to e-Learning Problems". Evolution of Teaching and Learning Paradigms in Intelligent Environment ISBN: 10.1007/978-3-540-71974-8_8 Volume 62, pp 183-221. Springer Berlin Heidelberg, 2007.
3. B.Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and, W. F. Punch."Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA" In Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE, 2003.
4. Talavera, L., and Gaudioso, E. "Mining student data to characterize similar behavior groups in unstructured collaboration spaces". In Proceedings of the Arti_cial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI ,Valencia, Spain, 2004.
5. Ş. Z. ERDOĞAN, M. TİMOR . "A data mining application in a student database". Journal of aeronautics and space technologies ,volume 2 number 2 (53-57) 2005.
6. G.J. Hwang. "A Knowledge-Based System as an Intelligent Learning Advisor on Computer Networks" Journal of Systems, Man, and Cybernetics Vol. 2 , pp.153-158, 1999.
7. G.J. Hwang, T.C.K. Huang,and C.R. Tseng. "A Group-Decision Approach for EvaluatingEducational Web Sites". Computers & Education Vol. 42 pp. 65-86 , 2004.
8. G.J. Hwang, C.R. Judy, C.H. Wu, C.M. Li and G.H. Hwang. "Development of an Intelligent Management System for Monitoring Educational Web Servers". In proceedings of the 10th Pacific Asia Conference on Information Systems, PACIS . 2334-2340, 2004.
9. G.D. Stathacopoulou, M. Grigoriadou. "Neural Network-Based Fuzzy Modeling of the Student in Intelligent Tutoring Systems". In proceedings of the International Joint Conference on Neural Networks. Washington ,3517-3521,1999.

10. C.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.): International Conference on Computational Science, ICCS 2001. LNCS Vol. 2074. Springer-Verlag, Berlin Heidelberg New York, 429-438. 2001.

11. S. Encheva , S. Tumin. “ Application of Association Rules for Efficient Learning Work-Flow” Intelligent Information Processing III , ISBN 978-0-387-44639-4, pp 499-504 published Springer Boston, 2007.

12. H.H. Hsu, C.H. Chen, W.P. Tai. “Towards Error-Free and Personalized Web-Based Courses”. In proceedings of the 17th International Conference on Advanced Information Networking and Applications, AINA’03. March 27-29, Xian, China, 99-104, 2003.

13. P. L. Hsu, R. Lai, C. C. Chiu, C. I. Hsu (2003) “The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance” [Expert Systems with Applications 25 (2003) 51–62.

14. A.Y.K. Chan, K.O. Chow, and K.S. Cheung. “Online Course Refinement through Association Rule Mining” Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 44, 2008.

15. S. Saxena, A. S.Pandya, R. Stone, S. R. and S. Hsu (2009) “Knowledge Discovery through Data Visualization of Drive Test Data” International Journal of Computer Science and Security (IJCSS), Volume (3): Issue (6) pp. 559-568.

16. S. Das and B. Saha (2009) “Data Quality Mining using Genetic Algorithm” International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (2) pp. 105-112

17. M.Anandhavalli , M.K.Ghose and K.Gauthaman(2009) “Mining Spatial Gene Expression Data Using Association Rules”. International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (5) pp. 351-357

18. R. Damaševicius. “Analysis of Academic Results for Informatics Course Improvement Using Association Rule Mining”. Information Systems Development Towards a Service Provision Society. ISBN 978-0-387-84809-9 (Print) 978-0-387-84810-5 (Online) pp 357-363, published by Springer US, 2009.

19. Ceglar, J.F Roddick. “Association mining”. ACM Computing Surveys, 38:2, pp. 1-42, 2006

20. S. Kotsiantis, , D. Kanellopoulos. “Association Rules Mining” A Recent Overview.GESTS Int. Transactions on Computer Science and Engineering, Vol. 32 (1), pp. 71-82, 2006.

21. H. Jochen, G. Ulrich and N. Gholamreza . **“**Algorithms for Association Rule Mining – A General Survey and Comparison”. SIGKDD Exploration, Vol.2, Issue 1, pp 58-64. ACM, 2000.