



Mining Students' Performance in Cisco Networking Academy Program Using Apriori Algorithm

Nenita D. Guerrero*, Shaneth C. Ambat
AMA University, Quezon City,
Philippines

Abstract— Recently, researchers and Higher Education Institutions (HEI) are also beginning to explore the potential of data mining in analyzing academic data. The goal of such endeavor is to find means to improve the services that these institutions provide and to enhance academic instructions. This type of data mining application is more popularly known as educational data mining or academic analytics. The study aims to explore the data mining technique using apriori algorithm in mining students' academic data related to students' performance in cisco subjects leading to cisco certification. The goal of the paper is to extract and to discover interesting relations between variables in the dataset and to identify strong rules discovered in databases. This scenario will help administrators identify which subjects or parameters are in need of in depth discussion and importance.

Keywords— apriori, data mining, academic data, association rules, analytics

I. INTRODUCTION

A. Background of the Study

Cisco Networking Academy Program (CNAP) courses are offered through blended learning that combines classroom instruction with online curricula, interactive tools, hands-on activities, and online assessments that provide immediate feedback. The instructor regularly meets students face-to-face supplemented by learning resources and activities available online.

The researcher conducted preliminary interviews in selected schools that offers CNAP. It was observed that many academy students lack confidence to pursue the certification exam. On the other hand, low percentage of certification passers have been identified compared to exam takers who failed the certification. Addressing this problem is essential, as universities and colleges offer courses as pre-requisites or subjects contributing to the development of students' mastery of network technologies. This scenario indicates the need to conduct research in this area in order to extract associated rules which will help the administrators to identify subjects or parameters which are in need of in depth discussion.

Research studies from HEIs already indicated that early identification of associated attributes in a set of parameters are keys to understanding students' academic behaviours. [1]-[3]

B. Research Questions

The study aims to address the following research questions:

1. What are the significant attributes that contribute in predicting students' success in cisco certification?
2. What association rules can be derived from the dataset?

II. LITERATURE REVIEW

A. Data Mining

Data Mining is an application of a specific algorithm in order to extract patterns from a database. KDD has become a very important process to convert this large wealth of data into business intelligence, as manual extraction of patterns has become seemingly impossible in the past few decades.

According to Rai (2014) [4] Data Mining (DM) is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repositories, etc.

B. Apriori Algorithm

Agrawal et al [5][6] discussed that one of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. The objective of Association rule mining is to find out the interesting rules from which knowledge can be derived.

The Apriori algorithm was the first attempt to mine association rules from a large dataset. The algorithm can be used for both, finding frequent patterns and also deriving association rules from them (Agrawal, 1994) [5].

C. Related Works

Association Rule Mining (Nuntawut et al., 2014) [9] is a technique that has been familiar in the relationship between the data that show most association rule mining in a variety of ways. Association mining that finds dependencies between the values of an attribute was introduced and has emerged as a famous research region.

According to Kotsiantis (2006) [10], there are two vital methods for association rules, support(s) and confidence(c). Since the database is huge and users worry about only those frequently purchased substance, regularly thresholds of support and confidence are determined by users to remove those rules that are not so interesting or needful. The two thresholds are known as minimal support and minimal confidence respectively.

III. METHOD OF ANALYSIS

The Knowledge Discovery in Databases [11] [12] was adapted and modified in a manner suitable for the development of this study. Fig. 1 below indicates the conceptual procedures conducted by the researcher.

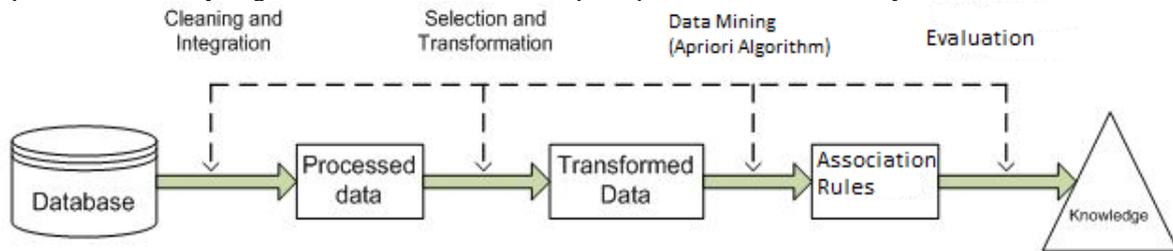


Fig. 1 Conceptual Model

The preliminary step involves selection of attributes needed to be performed based on goals. This section answers what data is needed and are these data available for consumption

A. Cleaning and Integration of Data

Data reliability is enhanced in this stage. It includes data clearing, such as handling missing values, and removing of outliers. The researcher used SQL script commands as a data processing technique to query and integrate relevant and potential data from databases.

B. Selection and Transformation of data

The data were transformed into a proper format that data mining tool can recognize. The generated comma separated values were converted into ARFF file – a file intended for WEKA for data processing.

The analysis of data was done using WEKA, formally called Waikato Environment for Knowledge Analysis. WEKA is an open source application that is freely available under the GNU general public license agreement.

C. Data Mining using Apriori Algorithm

Apriori algorithm is a data mining technique used to mine association rules from a database. Fayyad et al. (1996) [13] discussed that association rule mining is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

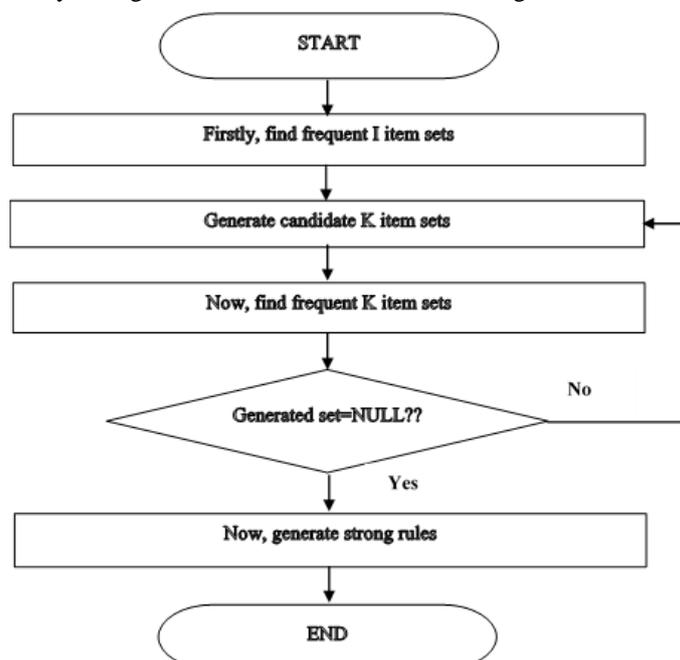


Fig. 2 Flowchart for Apriori Algorithm

Fig. 2 shows the flowchart of apriori algorithm according to Agrawal et al (1994) [6]

Nowadays, most association rule mining algorithms adapt the task decomposition strategy, which decomposes the association rule mining task into the two major subtasks [5] as follows:

- 1) Produce frequent itemset: find all itemsets that meet the minimum support threshold, called frequent itemset.
- 2) Produce rules: extraction all rules with high confidence from the produced frequent itemsets, called strong rule.

D. Evaluation

This step involves checking and resolving potential conflicts with the previously presumed knowledge.

IV. RESULTS AND DISCUSSION

A. Significant Attributes

To determine the statistical significance of a predictor the p -value was used. The predictor is statistically significant when a p -value is less than the significance level. The p -value is the probability of observing an effect given that the null hypothesis is true whereas the significance or alpha (α) level is the probability of rejecting the null hypothesis given that it is true. In practice, significance level is chosen before data collection and is usually set to 0.05.

This study used 18 potential attributes that may contribute to students' success in cisco certification as shown in Table 1.

Table 1. Potential Attributes To Predict Students' Success In Cisco Certification

Item	Attributes	Data Type
1	Gender	nominal
2	Scholarship	nominal
3	IT Electives	ordinal
4	IT Professional Subjects	ordinal
5	Core Subjects	ordinal
6	Social Science	ordinal
7	Cisco 1 Grade	ordinal
8	Cisco 1 Final Exam	ordinal
9	Cisco 1 Practical Exam	ordinal
10	Cisco 2 Grade	ordinal
11	Cisco 2 Final Exam	ordinal
12	Cisco 2 Practical Exam	ordinal
13	Cisco 3 Grade	ordinal
14	Cisco 3 Final Exam	ordinal
15	Cisco 3 Practical Exam	ordinal
16	Cisco 4 Grade	ordinal
17	Cisco 4 Final Exam	ordinal
18	Cisco 4 Practical Exam	ordinal

The model below was generated using SPSS. Results of the significant attributes processed using binary logistic regression is summarized in Table 2.

Table 2. Significant Attributes In Predicting Cisco Certification Using Regression Analysis

Attributes	B	S.E.	Wald	df	P value	Exp(B)	95% C.I.for EXP(β)	
							Lower	Upper
IT Professional Subjects	-0.959	0.375	6.533	1	0.011	0.383	0.184	0.800
IT Electives	1.109	0.326	11.577	1	0.001	3.031	1.600	5.742
Cisco 1 Grade	1.676	0.294	32.455	1	0.000	5.344	3.002	9.512
Cisco 3 Practical Exam	0.696	0.362	3.701	1	0.050	2.006	0.987	4.077
Cisco 4 Grade	0.973	0.303	10.297	1	0.001	2.646	1.460	4.793

Notes: * $p \leq 0.05$; the dependent variable in the analysis is cisco certification status coded so that 1=not pass and 2 = pass.

Analysis of the data reveals that five variables significantly predict students' performance in cisco certification; Information Technology elective subjects, cisco 1 grade, cisco 3 practical exam and cisco 4 have a positive β coefficient, indicating that the higher the scores of the students in the lists of predictors, the higher the likelihood that they will pass the cisco certification examination.

The coefficient reveals that IT Elective has a value of 1.109 coefficient where p value ≤ 0.05 , IT Professional - 0.959 where p value ≤ 0.05 Cisco 1 has a coefficient value of 1.679 where p value ≤ 0.05 , final practical cisco 3 has a coefficient value of 0.696 where p value < 0.05 and cisco 4 coefficient value 0.937 where p value ≤ 0.05 .

The positive β coefficients of the academic subjects indicate that the higher the grades of the students on such subjects the higher the odds of passing the certification.

B. Association Rules Generated from the Data set using Apriori Algorithm

The following are the best rules found in the data sets.

- 1) IT_Prof=2 Social_Science=2 123 ==> CoreProg=2 120
- 2) Cisco2=1 117 ==> Gender=1 114
- 3) Scholarship=1 IT_Prof=2 124 ==> CoreProg=2 12
- 4) Scholarship=1 IT_Prof=2 Certification=1 112 ==> CoreProg=2 108
- 5) Final_PracCisco4=2 115 ==> Scholarship=1 107
- 6) Final_PracCisco4=2 115 ==> Cisco2=2 107
- 7) Cisco1=1.5 155 ==> Certification=1 144
- 8) Final_ExamCisco4=2 119 ==> Cisco2=2 110
- 9) Gender=1 Cisco1=1.5 118 ==> Certification=1 109
- 10) CoreProg=2 IT_Prof=2 Certification=1 117 ==> Scholarship=1

The generated rules determine the strongest and associated attributes present in the data sets. The rules describe that most students who failed cisco examination are students who have lowest score in Cisco 1, this is in reference to the Cisco1=1.5 155 ==> Certification=1 with 144 data instances. The rule sets can reveal that students who have higher scores in cisco subjects can have the greater possibility in passing cisco certification examination

V. CONCLUSION

The study reveals that from the lists of potential attributes, Information Technology elective and professional subjects, cisco 1 grade, cisco 3 practical exam, and cisco 4 are the significant attributes in students' success for Cisco Certification. The study identifies that students who have higher scores in cisco subjects can have the greater probability in passing the cisco certification examination.

REFERENCES

- [1] Arnold, K. E., & Pistilli, M. D. (2012). *Course Signals at Purdue: Using Learning Analytics to Increase Student Success*. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 267-270). ACM.
- [2] Baeppler, P., & Murdoch, C. J. (2010). *Academic Analytics and Data Mining in Higher Education*. International Journal for the Scholarship of Teaching and Learning, 4(2), 17.
- [3] Campbell, J. P., & Oblinger, D. G. (2007). *Academic Analytics*. Educause Article.
- [4] Rai, P.K. (2014). *Descriptive Data Mining for Educational Environment*. International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 10, October 2014 ISSN: 2277 128X
- [5] Agrawal, R., Imielinski, T., & Swami, A. N. (1993). *Mining Association Rules Between Sets of Items in Large Databases*. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207-216.
- [6] Agrawal, R. & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 487-499.
- [7] Agrawal, R. & Srikant, R. (1995). *Mining sequential patterns*. In Eleventh International Conference on Data Engineering, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, 3-14.
- [8] Neelima, S., et al. (2015). *Enhanced Apriori Algorithm Using Top down Approach*. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 6, June 2015 ISSN: 2277 128X
- [9] Nuntawut, K., et al (2014). *A Technique to Association Rule Mining on Multiple Datasets*. Journal of Advances in Information Technology, vol. 5, no. 2, may 2014.
- [10] Kotsiantis, S. & Kanellopoulos, D. (2006). *Association Rules Mining: A Recent Overview*. GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82
- [11] Piatetski, G., & Frawley, W. (1991). *Knowledge discovery in databases*. MIT press.
- [12] Fayyad, U. M. (1996). *Data mining and knowledge discovery: Making sense out of data*. IEEE Expert: Intelligent Systems and Their Applications, 11(5), 20-25.
- [13] Fayyad, U.M. & Piatetsky-Shapiro, G. (1996). *Advances in Knowledge Discovery and Data Mining*. Editors: Usama M. Fayyad & Gregory Piatetsky-Shapiro, Cambridge, AAAI/MIT press, Pp. 1-625.
- [14] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. I. (1994). *Finding Interesting Rules from Large Sets of Discovered Association Rules*. In Proceedings of the third international conference on Information and knowledge management (pp. 401-407). ACM.