Predicting students' final passing results using the Apriori Algorithm

by Julianti Kasih Mewati Ayub & Sani Susanto

Submission date: 12-Aug-2021 01:08PM (UTC+0700)
Submission ID: 1630525644
File name: 010_Predicting_students_2.pdf (654.93K)
Word count: 2159
Character count: 11584

World Transactions on Engineering and Technology Education
Vol.11, No.4, 2013

Predicting students' final passing results using the Apriori Algorithm

Julianti Kasih†, Mewati Ayub† & Sani Susanto‡

Maranatha Christian University, Bandung, Indonesia†
Parahyangan Catholic University, Bandung, Indonesia‡

ABSTRACT: The studies discussed form part of a programme, other aspects of which have been previously considered [1][2]. The ultimate objective is to facilitate a lecturer in helping students to predict their final passing results based on their performance in several subjects in the first four semesters of their study period. In previous research, this aim was achieved through two techniques: discriminant analysis [1] and the Classification and Regression Trees (CART) algorithm [2]. Those two techniques resulted in a diagramme-based relationship. In this research, a rule-based relationship of the form *IF - THEN* is introduced and subsequently applied using software based on the Apriori Algorithm.

INTRODUCTION

This research project continues a theme that was considered in previous studies [1][2], the objective of which was to facilitate a lecturer in helping students to predict their final passing results based on their performance in several subjects in the first four semesters during their study period. The arguments for why this kind of prediction is considered important were discussed in [2] and are rewritten in the Appendix. The passing results in the Indonesian education system are classified into three grades: Extraordinary (Cum Laude), Very Satisfactory and Satisfactory [3].

The research was undertaken in the same institution, the Faculty of Information Technology, a university in Bandung, West Java, Indonesia. For reasons of confidentiality, the full name of the institution has not been included. In the two previous works, it was demonstrated that discriminant analysis [1] and the Classification and Regression Trees (CART) algorithm [2] helped academic advisors in this faculty to predict the final passing results of a student based on his/her grade in some subjects during the first four semesters during their undergraduate programme. This sort of facility enables academic advisors to assist students in setting up their study plans each semester in order for them to perform to their full potential [1][2]. Moreover, this work aims at helping the academic advisors with a more practical way of predicting the final passing results of a student.

In this research, a data mining task called an *association* was employed. Association is performed through a technique called the Apriori Algorithm. This algorithm produces some rule-based relationships in the form *IF- THEN* statements. This kind of statement serves in a more *ready to read* feature compared to the territorial map or decision tree employed in the previous work in [1] and [2], respectively.

OVERVIEW OF BACKGROUND THEORY

David Hand et al define data mining as *the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner* [4]. The observational data or the data to be summarised are often called the *training data.* Data mining has six tasks: description, estimation, prediction, classification, clustering and association [4]. Association is based on affinity analysis, the study of attributes or characteristics that *go together*. One amongst several methods for affinity analysis is market basket analysis, which tries to discover associations among these attributes with the aim to discover association rules *for quantifying the relationship between two or more attributes.*

The association rule takes the form *If antecedent, then consequent*, which for reasons of simplicity often desires a single consequent [4]. The performance measures of this rule are the *support, confidence, rule support, lift and deployability* outcomes. With the assistance of the SPSS Clementine 10.1 software package, these measures are first defined by the

term *instances*. Instances define the number of records in the data set that match the antecedents. For example, given the association *If purchase bread*, then *purchase cheese*, the number of records in the training data that include the antecedent *purchase bread* are referred to as instances.

Support or antecedent support is the proportion of training data for which the antecedents are true. For example, if 50% of the training data includes the purchase of bread, then, the rule *If purchase bread*, then, *purchase cheese* will have an antecedent support of 50%. Support as defined here is the same as the instances but is represented as a percentage.

Rule support is the proportion of training data for which the entire rule, antecedents and consequent(s), are true. For example, if 20% of the training data contains both the purchase of bread and cheese, then, rule support for the rule *If purchase bread*, then, *purchase cheese* is 20%.

Confidence is the ratio of rule support to antecedent support. This indicates the proportion of training data with the specified antecedent(s) for which the consequent(s) is/are also true. For example, if 50% of the training data contains bread (indicating antecedent support) but only 20% contains both bread and cheese (indicating rule support), then, confidence for the rule *If purchase bread*, then, *purchase cheese* would be rule support/antecedent support or, in this case, 40%. Lift is the ratio of confidence for the rule to the prior probability of having the consequent. For example, if 10% of the entire population purchases bread, then, a rule that predicts whether people will purchase bread with 20% confidence will have a lift of 20/10 = 2. If another rule tells that people will purchase bread with 11% confidence, then, the rule has a lift of close to one (1), meaning that having the antecedent(s) does not make a lot of difference in the probability of having the consequent. In general, rules with lift far from 1 will be more interesting, than, rules with lift close to one (1).

Deployability is a measure of what percentage of the training data satisfies the conditions of the antecedent but does not satisfy the consequent. In product purchase terms, it basically means what percentage of the total customer base owns (or has purchased) the antecedent(s) but has not yet purchased the consequent.

EXPERIMENT: THE RESULT AND INTERPRETATION

The research was undertaken in the same institution, the Faculty of Information Technology, a university in Bandung, West Java, Indonesia. The rules were generated by SPSS Clementine 10.1 software.

As in the previous research programme, the academic transcripts from 146 alumni served as input or observational or training data, which were available from the authors [1][2]. From these data, the students' *final passing results* were determined by the *final marks* from the following eight (8) subjects: IF102 (Introduction to Computer Application), IF103 (Introduction to Information Technology), IF104 (Algorithms and Programming), IF105 (Basic Programming), IF106 (Informatics Mathematics) and IF202 (Linear Algebra and Matrices), IF 203 (Computer Network) and IF 205 (File System and Access). The final marks of these eight subjects take the role of *antecedents*. The final marks of a subject were classified into five (5) groups, as follows: A (High Distinction), B (Distinction), C (Credit), D (Pass) and E (Fail) with some intermediates, such as B+ and C+.

The students' final passing results, as mentioned previously, were classified into three groups: 1 - Extraordinary (Cum Laude); 2 - Very Satisfactory; and 3 - Satisfactory. The students' final passing results take the role as the *consequent.*

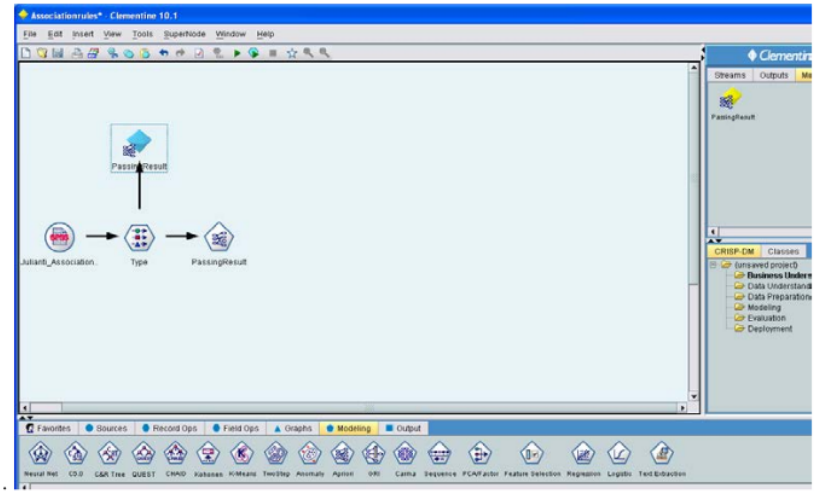

Figure 1: The classification model representation in Clementine 10.1.

The SPSS Clementine 10.1 model is described in Figure 1. In this figure, the icon on:

- the left side describes the input or observational or training data;
- the right side describes the algorithm employed, in this case, the Apriori Algorithm; and
- the top-middle side describes the output of the Apriori Algorithm;
- the bottom-middle side describes the type of the data, the final marks of the 8 (eight) subjects are the antecedents, while the passing result is the consequent.

The training data were saved in the form of a SPSS worksheet file, a section of which is represented in Figure 2.

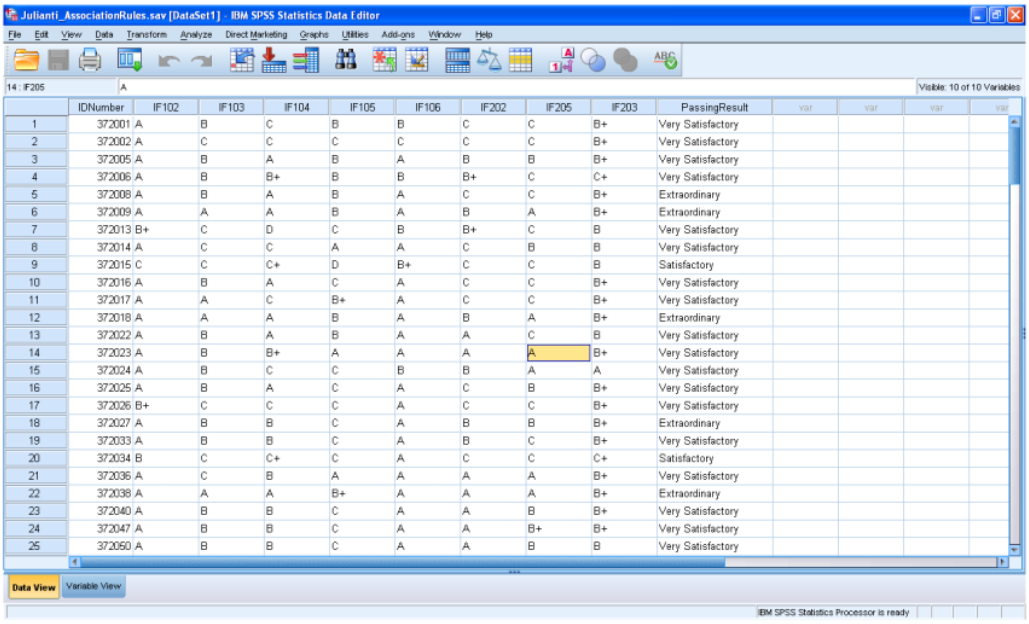

Figure 2: Some part of the training data.

The generated association rules are displayed in Figure 3. In generating association rules, SPSS Clementine 10.1 software gives the user the options to determine the minimum antecedent support, the minimum rule confidence and the maximum number of antecedents, which in this research were set at 20%, 80% and 5, respectively. Six association rules were generated. The authors will be able to generate more rules if they reduce the minimum antecedent support and the minimum rule confidence and *vice versa*.

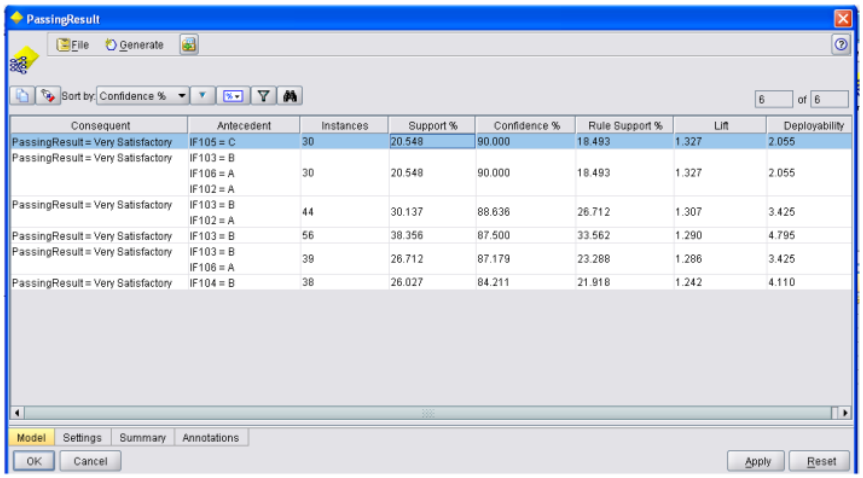

Table 3: The generated association rules.

One of the rules generated is *IF IF103 = B and IF106 = A and IF102 = A, THEN Passing Result = Very Satisfactory*. The antecedents of this rule is *IF103 = B and IF106 = A and IF102 = A*, and its consequent is *Passing Result = Very Satisfactory*. This rule has the following performances:

- Instances equal to 30, which means that out of 146 records in the training data sets, the number of records in the data set that match the antecedents is 30;
- Support or antecedent support is 20.548%, this value is due to the fact that the number of records in the 146 training data for which the antecedents are true is 30, or 20.548%;
- Rule support is 18.493%, this value is due to the fact that the number of records in the training data for which the antecedents and consequent(s) are true is 27 out of 146 records or 18.493%;
- Confidence is 90%, which means that the ratio of rule support, that is 18.493%, to antecedent support, that is 20.548% is 90%;
- Lift is 1.327, this value comes from the ratio of confidence for the rule (90%) to the prior probability of having the consequent *Passing Result = Very Satisfactory* (99 records out of 146, or 67.8%);
- Deployability is 2.055%, this value comes from a measure of what percentage of the training data satisfies the conditions of the antecedent *IF103 = B and IF106 = A and IF102 = A*, but does not satisfy the consequent *Passing Result = Very Satisfactory*. In this case is three (3) records out of 146 or 2.055%.

Care should be taken when applying the generated rules. Those rules do not express the causal relationship between the antecedent(s) and the consequent. As discussed in the background theory, association is based on affinity analysis, the study of attributes that *go together*. In the association rule context, the attributes are the antecedent(s) and consequent.

CONCLUSIONS AND SUGGESTION FOR FURTHER RESEARCH

This research demonstrates the ability of one of the data mining techniques that enable an academic advisor to help students predict their final passing results. Compared to the previous research [1][2], this research demonstrates that the prediction can be performed very practically, with neither graphs nor charts being required. The prediction is carried out by sentences of the form, *If antecedent, then consequent*. Further research may take the form of the investigation of the prediction of the final passing results, which might be based on a multivariate statistics technique called *logistic regression*. The authors intend to perform this research in the not too distant future.

ACKNOWLEDGMENT

Sincerest thanks are conveyed by the authors to Mr Radiant Victor Imbar, former Dean of the Faculty of Information Technology at Maranatha Christian University, for his support and providing the data from the academic transcripts of alumni.

REFERENCES

1. Kasih, J., and Susanto, S., Predicting students' final results through discriminant analysis. *World Transactions on Engng. and Technol. Educ.*, 10, **2**, 144-147 (2012).
2. Kasih, J., Ayub, M. and Susanto, S., Predicting students' final passing results using the Classification and Regression Trees (CART) algorithm. *World Transactions on Engng. and Technol. Educ.*, 11, **1**, 46-49 (2013).
3. Government Regulation, the Republic of Indonesia, Nr 60 Year 1999 about Higher Education.
4. Larose, D.T., *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. (2005).

APPENDIX

In relation to the Introduction, the arguments for making the prediction of students' final passing results as the main issue of this research are as follows:

- First, one of the important aims of higher education in the Republic of Indonesia is to prepare the academic participants (students) to become members of society with the academic and/or professional abilities to enable them to apply/develop/enrich the foundations of knowledge in the sciences, technology and the arts.
- Second, to achieve this aim, undergraduate students are assigned to several academic advisors (an informal translation of the Indonesian term *Dosen Wali*) throughout their years of higher educational studies.
- Third, academic advisors, who are lecturers, have as their main task the fostering of students' academic and non-academic activities. With regard to students' academic activities, one of the duties of the academic advisor is to help students in setting up their study plans for each semester.
- Fourth, setting up a study plan includes providing guidance for students regarding how many subjects, and which subjects, to undertake.
- Fifth, through this guidance, students are expected to obtain the best passing results at the end of their undergraduate study [1-3].

Predicting students' final passing results using the Apriori Algorithm

ORIGINALITY REPORT

15%	9%	16%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	epdf.pub Internet Source	2%
2	mafiadoc.com Internet Source	2%
3	Huang Changhai, Hu Shenping. "Factors correlation mining on maritime accidents database using association rule learning algorithm", Cluster Computing, 2018 Publication	2%
4	Submitted to Indian Institute of Management, Bangalore Student Paper	2%
5	Peter R. Bakhit, BeiBei Guo, Sherif Ishak. "Crash and Near-Crash Risk Assessment of Distracted Driving and Engagement in Secondary Tasks: A Naturalistic Driving Study", Transportation Research Record: Journal of the Transportation Research Board, 2018 Publication	2%

6	Submitted to Grand Canyon University Student Paper	2%
7	Leon, Carlos, Félix Biscarri, Iñigo Monedero, Juan Ignacio Guerrero, Jesús Biscarri, and Rocío Millan. "Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies", IEEE Transactions on Power Systems, 2011. Publication	1%
8	Larose, . "Association Rules", Discovering Knowledge in Data, 2014. Publication	1%
9	Seyed Alireza Samerei, Kayvan Aghabayk, Amin Mohammadi, Nirajan Shiwakoti. "Data mining approach to model bus crash severity in Australia", Journal of Safety Research, 2020 Publication	1%
10	Siyi Yu, Jie Yang, Mingxiao Yang, Yan Gao, Jiao Chen, Yulan Ren, Leixiao Zhang, Liang Chen, Fanrong Liang, Youping Hu. "Application of Acupoints and Meridians for the Treatment of Primary Dysmenorrhea: A Data Mining-Based Literature Study", Evidence-Based Complementary and Alternative Medicine, 2015 Publication	1%
11	www.billytan.net Internet Source	

1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		