Examine study pattern on selective cross join data using bootstrap

Eka Suswaini[1*], Budi Warsito [2], Adi Wibowo[3]

{ekasuswaini@students.undip.ac.id [1], budiwarsito@lecturer.undip.ac.id [2], adiwibowo@lecturer.undip.ac.id[3]}
Universitas Maritim Raja Ali Haji[1], Universitas Diponegoro, Semarang[1,2,3]

Abstract. A learning analytics model uses students' academic records to recommend study paths based on the their academic performance. It also encourages students to improve their performance on the subjects in which they had a lower grade. Subsequently, the process of implementing a learning analytic system for study path recommendation can be carried out by developing a knowledge base model using selected cross-join data. In this study, the selective cross-join technique, which was implemented using the bootstrap validation method, was examined. Furthermore, the data used are drawn from student records from the previous two academic years that have already undergone pre-processing to eliminate any newly added courses, since there would not be much to learn from them. The validation process, which took 10 iterations, was carried out using the bootstrap method and the result for each iteration was evaluated using 1 - Root Mean Square Error. The lowest, highest, and average accuracies obtained from all 10 iterations were 69.2%, 92.3%, and 84.69%, respectively. This inconsistency indicated that the process may have been misinterpreted without taking into account any noise that might have been replicated in the data.

Keywords: Learning Analytics, Educational Data Mining, Selective Cross Join, Bootstrap validation.

1 Introduction

The primary methodology examined in this study is comparable to that of Matulatan and Resha, who used theirs to create the learning analytics model for all students by providing each course with a connection to the subsequent subject they will be taking in the upcoming semester. However, their research method had some performance and accuracy problems. In this paper, the accuracy of the results was investigated by creating a bootstrap dataset. In previous research, it was noted that the teachings of pedagogical practices, which involves collaborations and interactions, such as constructivist learning, have produced more significant result in terms of instilling the knowledge of general skills into university students [2]. This method merely aims to enhance the student's ability to learn the skill. However, the core notion of this paper is not the learning process but rather the potentials of making predictions using prior knowledge. Learning analytics has been an emerging topic over the last 5 years, and this involves the use of computational approaches to analyze available data [3][4]. As previously stated, the majority of the studies are focused on teaching methods alone. In this study, however, selective cross-join data

ICSEDTI 2022, October 11-13, Tanjungpinang, Indonesia

DOI 10.4108/eai.11-10-2022.2326328

was used as an alternative to help academic advisors analyze and recommend study paths to students based on their performances. In order words, the students' past performances are used to predict future results. Also, the bootstrap method was employed to create more combination data in order to get more accurate measurements.

2 Boundaries and Limitation

Before implementing the method of selective cross join on academic records, lets discuss the environment used in this study:

- The study taken place in Indonesia. Academic system used in this study will be Indonesia Higher Education System.
- The grade system will follow the institution academic grade rules.
- Any changes in curriculum courses could lead to new information that never exist before. This new information will be orphaned data (no pair).
- Any new courses that has been recently introduce will be sorted out since it will have not much information

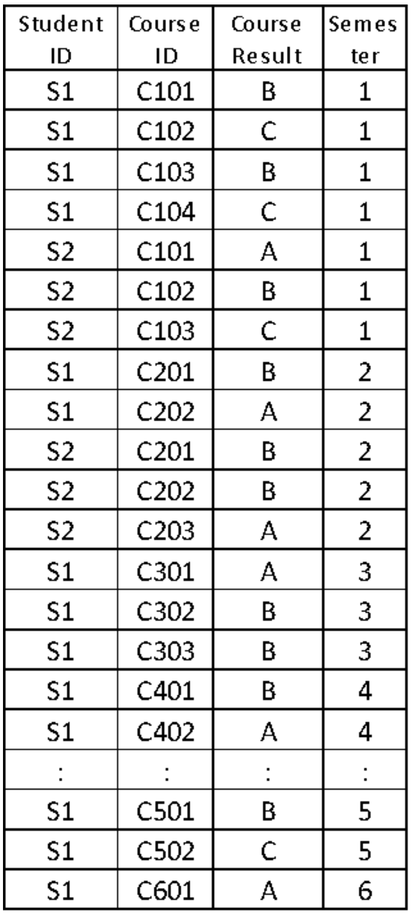
The limitation on this study are:

- CPU Cost (Time of the whole process) is not taking into consideration
- Assuming that there is no subjective grading happened.

3 The process

The process as describe in the source paper as following:

The process start with selection of the required attributes of the record (in this study, we are using student’s reference ID, course’s reference ID, semester (odd or even) and course grade result), collected it on one table (fig.1),



Student ID	Course ID	Course Result	Semester
S1	C101	B	1
S1	C102	C	1
S1	C103	B	1
S1	C104	C	1
S2	C101	A	1
S2	C102	B	1
S2	C103	C	1
S1	C201	B	2
S1	C202	A	2
S2	C201	B	2
S2	C202	B	2
S2	C203	A	2
S1	C301	A	3
S1	C302	B	3
S1	C303	B	3
S1	C401	B	4
S1	C402	A	4
:	:	:	:
S1	C501	B	5
S1	C502	C	5
S1	C601	A	6

Fig 1. Example of data arrangement in one table

This dataset would be used in bootstrap iteration. Because of the random selection in bootstrap process, then it could not be determined how the outlook of the bootstrap dataset result.

With each iteration, the bootstrap create new dataset, the next step is pairing every data row from semester i to semester i+1 using selective cross join, which is applying certain rule that could data could be joined. (fig 2). Any connection has weight from number of occurance, i.e, number of students who get A+ on course C101 then get A+ on course 201.:

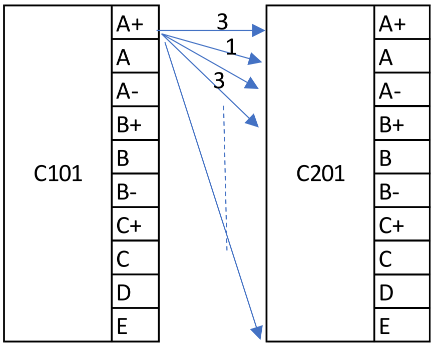


Fig 2. Cross join every possible grade from course semester i with semester i+1

Then calculate representative weight using correlation statistic like pearson to show most likely occurs (increase or decreased grade or not change)

The representative weight for overall connection would created any possible tree rute and data from the last semester will be the leaf or end node. The sorter the rute much be considered as prefered scenarios. (fig 3)

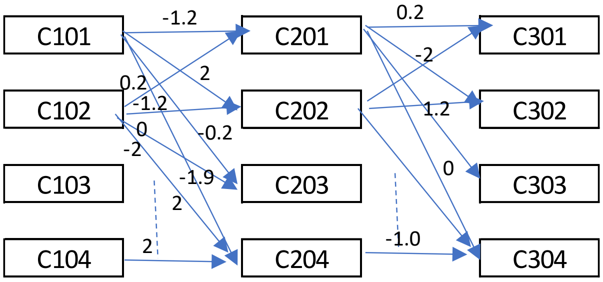


Fig 3. Fully Connected representative weight wirh layers represent semester

The last step is creating the model function that in original paper using Upper confidence Bound , but due to consideration on heavy calculation, the function will be replaced with simple bounding function that selecting path with better weight performance (i.e highest positive value)

4 Analysis

After the model had been built, the validation process was then implemented using bootstrap. Furthermore, in order to ensure that no two iterations will produce the same sequence number, a dummy training dataset that is the same size as the original dataset was first created. After this, the training dataset was filled with data from the original dataset using a random seed timer.

As aforementioned, the validation process was iterated ten times. In each iteration, the original dataset's leftover data was used for testing the model after which the Root Mean Squared Error was employed to demonstrate how accurately the model would describe each data testing scenario.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{n}} \dots \quad (1)$$

The finding is in Table 1

Table 1 Testing results of the 10 iterations

Iteration	Number of data left over	RMSE
1	13	0.85
2	45	0.76
3	36	0.86
4	40	0.69
5	24	0.92
6	38	0.81
7	13	0.86
8	38	0.90
9	11	0.92
10	16	0.90

The obtained average Root Mean Square Error of all ten iterations was 0.8469 or 84.69%. The accuracy of the distribution is shown in Fig 4.

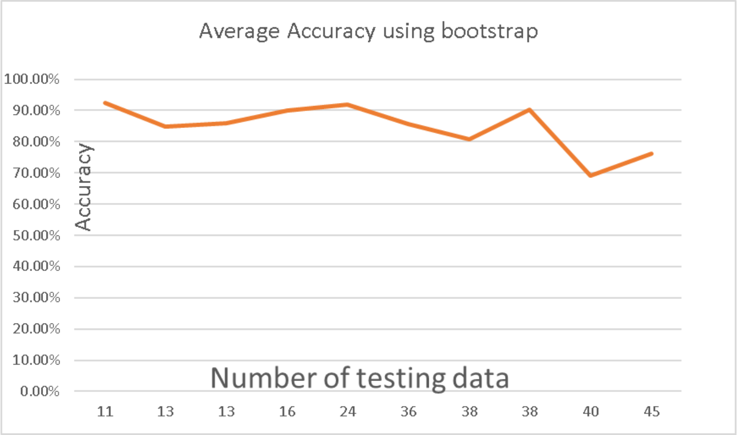


Fig 4 Distribution of Accuracy on number of testing data

5 Conclusion

In conclusion, the model's instability during bootstrap validation indicated that the method is flawed, and this results in inconsistent behavior. This inconsistency is triggered by scenarios where students fail a particular subject more than once and these instances represent the outliers in the dataset. Furthermore, the time taken for the whole process to complete is dependent on the leftover data volume that was used in testing. The total procedure was lengthy and had an unfavorable complexity of O(2n).

Acknowledgements

The data collected from private source, used with respect to privacy act.

References

[1] Buckingham, S. S., Deakin, R. C. : Learning Analytics for 21st Century Competencies, J. Learn. Analytic. 3(2) pp. 6–21 (2016)

[2] Mangaroska K., Giannakos, M. : Learning Analytics for Learning Design: A Systematic Literature Review of Analytics-Driven Design to Enhance Learning, IEEE Trans. Learn. Technol. 12(4) pp. 516–534 (2019)

[3] Brown, M., DeMonbrun, R M.,Teasley S. :Taken together: Conceptualizing students' concurrent course enrollment across the post-secondary curriculum using temporal analytics. Journal of Learning Analytics, 5(3), pp. 60–72 (2018)

[4] Hilliger I., Aguirre C., Miranda C., Celis S., Pérez-Sanagustín, M.: Design of a curriculum analytics tool to support continuous improvement processes in higher education. Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), 23–27 March 2020, Frankfurt, Germany , pp. 181–186. ACM Press (2020).

[5] Gottipati, S., Shankararaman, V.:. Competency analytics tool: Analyzing curriculum using course competencies. Education and Information Technologies, 23(1), pp. 41–60 (2017)

[6] Chou C-Y, Tseng S-F, Chih W-C, Chen Z-H, Chao P-Y, Lai K R, Chan C-L, Yu L-C. , Lin Y-L: Open student models of core competencies at the curriculum level: Using learning analytics for student reflection. IEEE Transactions on Emerging Topics in Computing, 5(1), 32–44, (2017)

[7] Gonzalez-Brenes, J. P., & Huang, Y.:. Using data from real and simulated learners to evaluate adaptive tutoring systems. In J. Boticario & K. Muldner (Eds.), Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education (AIED 2015), 22–26 June 2015, Madrid, Spain (**5**), pp. 31–34 (2015)

[8] T. Matulatan and M. Resha, "Deep learning on curriculum study pattern by selective cross join in advising students ' study path."

[9] A. Virtanen and P. Tynjälä, "Factors explaining the learning of generic skills: a study of university students' experiences," Teach. High. Educ., vol. 24, no. 7, pp. 880–894, Oct. 2019.

[10] S. Buckingham Shum and R. Deakin Crick, "Learning Analytics for 21st Century Competencies," J. Learn. Anal., vol. 3, no. 2, pp. 6–21, Sep. 2016.

[11] K. Mangaroska and M. Giannakos, "Learning Analytics for Learning Design: A Systematic Literature Review of Analytics-Driven Design to Enhance Learning," IEEE Trans. Learn. Technol., vol. 12, no. 4, pp. 516–534, Oct. 2019.

[12] McEneaney J ., Morsink,P. : Curriculum Modelling and Learner Simulation as a Tool in Curriculum (Re)Design. Journal of Learning Analytics, 9(2) pp 161-178 (2022).