IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 001, LIR005
DOI: https://doi.org/10.52731/lir.v001.005

Progressive Prediction of Students' Future Performance on Coherent Vertical Curriculum

Satrio Adi Priyambada [*], Fatharani Wafda [*],
Tsuyoshi Usagawa [*], Mahendrawathi ER [†]

Abstract

Predicting students' future performance is important for academic stakeholders as the students' success is the objective of the higher educational institutes. The prediction based on past performance and alignment with the curriculum is crucial to support decision-making action effectively for the university with a coherent vertical curriculum. The result of the prediction can be used to intervene and ensure that the student can graduate on time, also preventing the student from dropping out. In this paper, we proposed a methodology for predicting progressively the students' future assessment including feature engineering. Using the ensemble learning techniques, we adapt the existing Ensemble-based Progressive Prediction so it can be applied on students' data that used the Coherent Vertical Curriculum. In this paper, the behavioral data is used instead of domain knowledge-based data. The results show that the algorithm's accuracy has been improved on a real-world student dataset.

Keywords: educational data mining, ensemble learning, learning behavior, students' performance.

1 Introduction

Digitalization of the academic process in higher educational institutes generates a huge amount of academic data. This collection of data including the records of students' activity, grades, and other information can be used by academic stakeholders to gain insights for decision-making. The decision-making action can be an intervention to increase student graduation rate and prevent drop out. Moreover, obtaining knowledge from the academic data can be used to improve the curriculum guidelines.

The application of data mining techniques in educational domain can be referred as Educational Data Mining (EDM) [1]. There have been research in this area that provides the method to predict students' future performance in order to increase graduation rates and prevent drop out based on different aspects in the academic process. This kind of research uses the course level data or the curriculum level data which commonly stored in the E-learning or Learning Management Systems (LMS). The course level data are used to predict the grade of the students in a particular course, while the curriculum level data used the academic data such as course-taking activity or a whole student data from the beginning academic year to predict the Grade Point Average (GPA).

Different knowledge can be obtained from many perspectives. The research that focuses on educational data modeling and process analysis can be referred as Educational Process Mining (EPM) [2]. As part of EPM, curriculum mining aims to analyze the behavior of students on behalf of the curriculum, such as how they take the courses [3].

[*] Kumamoto University, Kumamoto, Japan
[†] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Understanding the learning behavior is one of the important things, especially for the university that used the coherent vertical curriculum. In this type of curriculum, the student needs to take the course exactly like in the curriculum guideline because there are academic standards to be achieved when student finished all course in the curriculum [4]. Indonesia is one of the countries that applied this type of curriculum to comply with the government guideline of standard competencies [5]. However, there are differences between students learning behavior with the curriculum guideline caused by various reason such as student failing the course, taking incoming courses in advanced, or credit limitation that prohibit student to take more course in particular semester.

The profile-based approach has been done by clustering the students per unit of time to understand students' learning behavior in comparison to the curriculum [5]. Also, the changes of the behavior from semester to semester have been analyzed using the migration of the students' learning behavior of the student at a certain period of time by clustering method based on students' profiles [6]. However, no clear results can be obtained by only analyzing the cluster of the students. Many research in EDM is utilizing prediction. By giving the estimated result of the students' future performance, it gives clearer idea to academic stakeholders in the decision making.

Predicting student performance in higher education requires continuous tracking and updating since the student completes new courses every certain period of time (semester in our case) [7]. The important thing is that the prediction of the students' performance is not a one-time prediction. Especially in coherent vertical curriculum, the other thing that needs to be considered for predicting the students' performance is the changing of students' learning behavior since each semester student take a new course whether the student will be passed or failed. The student might take a new course for the next semester or even retake the course when they failed. The ensemble progressive prediction exploit the result of prediction from previous time to take a portion on the result of the prediction from current time and make it will be progressively predicting the results.

In this paper, we aim to create prediction on students' performance, especially in the higher education institution that used coherent vertical curriculum. For this purpose, we use feature engineering to implement learning behavior which is the alignment between students course data and curriculum guideline.

2 Related Work

2.1 Educational Data Mining

The data mining techniques can be utilized as support on decision-making in the educational process. There has been research on creating a recommender system for the student to make a decision for selecting a subject by using the prediction on the performance and risk [8]. Also, the use of machine learning techniques in educational data can produce models that can be incorporated in decision-making process by the strategic level of the higher educational institution [9].

Several works have been conducted on analyzing the students' performance in the coherent vertical curriculum by using cluster analysis both on aggregated and segmented data [5], [10]. There is also research implementing the cluster evolution analysis to analyze the migration of the students' learning behavior [6]. However, it is difficult to get the insight effectively from this kind of results because it was used to monitor the performance of the student in current semester rather than predict the future performances. The purpose of the current work is to predict students' future performance based on students' past performance including the alignment with curriculum to create a better insight to help academic stakeholder in decision making action.

2.2 Student Performance Prediction

There have been many research that aim to predict the future performance of students by utilizing various resources such as the data from the Massive Open Online Course (MOOC) for student dropout prediction by using the neural network that can automatically extract the feature needed from raw data [11]. Classification can be used to optimize the prediction results. Research work in [12] used personal and academic data to predict the student performance and then classify students individually based on their talents.

In the [13], the interaction of the students and course have been used to predict student's grade and the author have found that there are relationship between students and students, courses and courses, students and courses. Behavior also can be exploited as feature in the prediction such as academic behavior, learning behavior, exam behavior [14], [15].

In the [7], Ensemble learning is utilized to get a better performance of the prediction. This work can be used to progressively predict students' performances by using both academic state such as SAT score and evolving data such as the credit and GPA. The ensemble predictor utilized the result of base prediction and previous ensemble prediction. The students' data used are typically credit and grade of cluster of the course base on its domain knowledge.

Our research will focus on predicting the students' future performance progressively by using the learning behavior data and the prediction result from previous semester by adopting the existing Ensemble Progressive Prediction (EPP) algorithm [7]. The vertical coherent curriculum is the curriculum that coherently structured based on the standard competencies from government. In this kind of curriculum, student cannot freely take the courses, because there is a set of course that they must take. However there are several situations that make the student take the course differently from curriculum guideline. Behavioral data can accommodate on how the student take the course in reality by using the alignment of the student's course registration data and curriculum guideline. The prediction will be executed every semester by utilizing students' past performance. However, instead of using the domain knowledge-based data, we use the behavioral data.

3 Methodology

To address the aforementioned problems, we propose a framework for predicting the performance of the students especially for the higher education that used the vertical coherent curriculum. We use the students' learning behavior based on the course-taking activity and predict the performance by using modified EPP.

3.1 Data Preprocessing

Data preparation aims to obtain students' features that will be used in each semester from the institution's database. In the Vertical Coherent Curriculum, student i must complete course strictly specified by curriculum K to graduate. The students who graduate on time is the students who graduates within 8 semesters ($T=8$) and does not take any course after 8th semester.

There are two type of feature that will be used in our case 1) base data; and 2) evolving data. Base data is the data that belong to only that time period which is semester in our case. Evolving data is the data that contains more information such as GPA and cumulative obtained credits.

Most of the data utilized in this work will be credit and grade of the course. We reduce the number of features by aggregating the data of the student for each semester by clustering the data. Let $d_i^t \in D^t$ denote student *i*'s base data at semester *t*={1,2,…}. Denote $d_i^t = B_i^t\ .\ GPAS_i^t$ as the *d* is a set of student *i*'s base feature in semester t, where B_i^t is a set of behavioral data of student, $GPAS_i^t$ is the grade point average semester.

The set of possible learning behavior denoted by *LB={m,a,b}* which is the learning behavior *match, after* and *before* respectively. The learning behavior of course-taking activity can be obtained by looking up the alignment between students' course and curriculum guideline [5]. The course taken in each semester aggregated by its credit and obtained grade by this students learning behavior. Denote $B_i^t = C_i^{LB}\ .G_i^{LB}$ as the set of learning behavior, C_i^{LB} is the total credit of the particular learning behavior, and G_i^{LB} is the total grade of the particular learning behavior of the student I in the semester *t*.

The Grade Point Average-Semester (GPAS) is defined as

$$GPAS^t = \frac{1}{C^t}\Sigma_j^t(C_j^t\ .\ G_j^t) \qquad (1)$$

where C_j^t is the credit and G_j^t is the grade of the *j*-th course in the *t*-th semester and C^t is the total credit of all courses in *t*-th semester.

Evolving data is the data that contain information not only from one point of time but also the cumulative of previous data. Let the $e_i^t \in E^t$ denote student *i*'s evolving data at semester *t*. Denote $e_i^t = B_i\ .GPA_i^t$ as the e is a set of student i's evolving feature in semester t, where B_i is a set of cumulative behavioral data of student and GPA_i^t is the grade point average.

Combined data is the combination of the base data and evolving data. Let the $r_i^t = d_i^t\ .e_i^t$ where r_i^t is a set of student *i*'s combined data in the semester *t*, d_i^t is a set of base data and e_i^t is a set of evolving data.

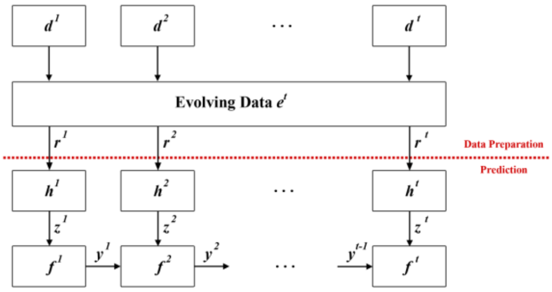

Figure 1: Prediction Methodology

3.2 Ensemble-based Progressive Prediction

To address the problem, we modified the existing EPP algorithm, so it can be used on the students' data with vertical coherent curriculum. We re-design the predictors so it can progressively predict the students' performances. The existing EPP would likely use the curriculum that student can take freely. In our case, the students mostly take the same course and can take about 3 courses for the elective course. Also, student who got good GPA Semester will have extra additional credits that can be an advantage so they can take extra courses from incoming semester ahead. On the other hand, student who got bad GPA Semester will have their credits limited in the next semester. In this situation, the important thing that need to be considered is how students will react to this condition. Thus, in this research, we use behavioral data instead of domain knowledge data.

We adopt the EPP to design predictors for progressively expanding input space. Given the base data and the evolving data, we construct a predictor $h^t : r^t \rightarrow z$ where z is the binary of whether student will graduate on time or not.

Our ensemble predictors using the weighted moving average which is designed as such that older result are given lower weights than current result. The ensemble prediction is

$$f^t : v.y^{t-1} + w.z^t \qquad (2)$$

where y is the binary of whether student will be graduate on time or not. For the student i in the semester t, the ensemble predictor f^t make a prediction y_i^t based on a weighted previous ensemble prediction y_i^{t-1} and weighted prediction result z_i^t. Weight vector v is associated with the result of ensemble prediction f^t and weight vector w is associated with the result of prediction h^t. Because of the evolving prediction, the current prediction results have more weight than previous ensemble prediction, so the weighted vector $w > v$ and $w + v = 1$. The diagram for the methodology that consist of the data preparation and the prediction phase is illustrated in Figure 1.

4 Experiments

4.1 Dataset

The student data used is collected from an Academic Information Systems of an Information Systems Department. The dataset has 635 anonymized students enrolled with 61% of the students graduated on-time.

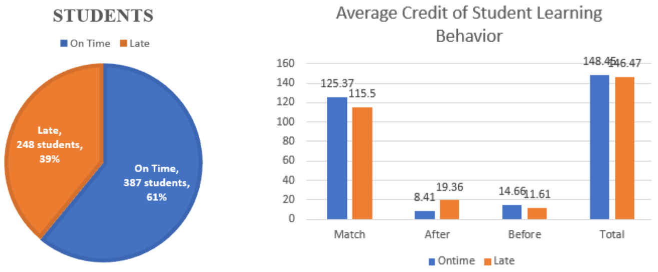

Figure 2: Students' Data (left) and Average Credit of Student Learning Behavior

The data of the students consist of the students' GPAS, GPA, the courses that the students take in each semester, the credits and the obtained grades. Because the prediction is not a one-time prediction, we divide the base data d^t by 8 which is the number of the semester to normally graduated for bachelor degree in our case. We create the evolving data e^t from each of the base data per semester. Also we combine r^t of each base and evolving data for each semester that will be used for the prediction. Students who graduated on time had a higher GPA and obtained more credits up to 8th semester. This due to the students who graduated late still had unfinished courses that must be taken after 8th semester. Also from this information we can see that students who graduated on time take more *match* course and tend to more align with the curriculum guideline. The data consist of 6 batch of students with 3 different curriculum due to the change of curriculum every five years. The average total number of course is 54 courses with 42 course are compulsory courses and only 9 credits need to be taken from 12 elective course choices.

4.2 Performance

For the performance result of EPP algorithm for vertical coherent curriculum, we compare the performance of the algorithm with:

- Only using the base predictor d_i^t, which consist of only the data of student *i* for *t* semester.
- Only using the evolving predictor e_i^t, which consist of the cumulative data of student *i* up to semester *t*.
- Use combined predictor r_i^t. We implemented several algorithms such as decision tree, logistic regression, and K-Nearest Neighbour (KNN) algorithm to predict the data, however KNN yields the highest average accuracy over all the semesters. Thus, we only show the result of the KNN.
-

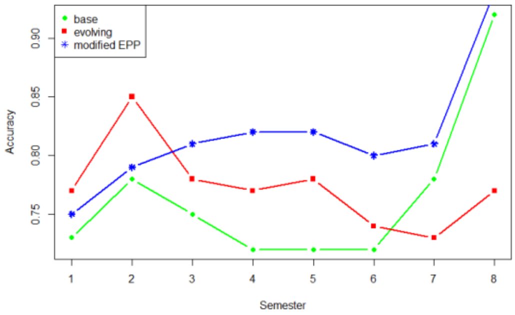

Figure 3: Prediction Accuracy

From Figure 3, as we can see that modified EPP outperformed the other at semester 3-8 in the prediction. The prediction accuracy of evolving data had bigger value in the first two semester, however the accuracy declined after the second semester. The details of the prediction accuracy comparison can be seen on Table 1.

Table 1: Prediction Accuracy

Semester	1	2	3	4	5	6	7	8
Base	0.73	0.78	0.75	0.72	0.72	0.72	0.78	0.92
Evolving	0.77	0.85	0.78	0.77	0.78	0.74	0.73	0.77
Modified EPP	0.75	0.79	0.81	0.82	0.82	0.8	0.81	0.94

5 Conclusion

In this paper, we modified the Ensemble-based Progressive Prediction so it can be used in students' data that used Vertical Coherent Curriculum. This curriculum has strict rule that needs to be considered by students. Despite of this, students still had unique and diverse data that can be seen by using the learning behavior of the course-taking activity.

This approach can be used for evaluating students' performance whether she/he will be graduate on time or not and provide the information for academic stakeholders to make a decision that can support students.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research 19H1225100.

References

[1] R. S. J. d. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *J. Educ. Data Min.*, vol. 1, no. 1 SE-Articles, pp. 3–17, Oct. 2009, doi: 10.5281/zenodo.3554657.

[2] N. Trčka and M. Pechenizkiy, "From local patterns to global models: Towards domain driven educational process mining," *ISDA 2009 - 9th Int. Conf. Intell. Syst. Des. Appl.*, pp. 1114–1119, 2009, doi: 10.1109/ISDA.2009.159.

[3] M. Pechenizkiy, N. Trcka, P. De Bra, and P. Toledo, "CurriM : Curriculum Mining," *Proc. 5th Int. Conf. Educ. Data Min.*, no. i, pp. 1–4, 2012, [Online]. Available: http://educationaldatamining.org/EDM2012/uploads/procs/Posters/edm2012_poster_1 1.pdf%5Cnhttp://www.win.tue.nl/~mpechen/projects/edm/CurriM_extract.pdf.

[4] National Research Council, *Systems for State Science Assessment*. Washington, D.C.: National Academies Press, 2005.

[5] S. A. Priyambada, M. Er, and B. N. Yahya, "Curriculum Assessment of Higher Educational Institution using Segmented-trace Clustering," *J. Tek. Ind.*, vol. 20, no. 1, pp. 33–48, 2018, doi: 10.9744/jti.20.1.33-48.

[6] S. A. Priyambada, M. Er, B. N. Yahya, and T. Usagawa, "Profile-Based Cluster Evolution Analysis: Identification of Migration Patterns for Understanding Student Learning Behavior," *IEEE Access*, vol. 9, pp. 101718–101728, 2021, doi: 10.1109/ACCESS.2021.3095958.

[7] J. Xu, Y. Han, D. Marcu, and M. Van Der Schaar, "Progressive prediction of student performance in college programs," *31st AAAI Conf. Artif. Intell. AAAI 2017*, pp. 1604–1610, 2017.

[8] A. J. Fernández-García, R. Rodríguez-Echeverría, J. C. Preciado, J. M. Conejero Manzano, and F. Sánchez-Figueroa, "Creating a recommender system to support higher education students in the subject enrollment decision," *IEEE Access*, vol. 8, pp. 189069–189088, 2020, doi: 10.1109/ACCESS.2020.3031572.

[9] Y. Nieto, V. Gacia-Diaz, C. Montenegro, C. C. Gonzalez, and R. Gonzalez Crespo, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 7, pp. 75007–75017, 2019, doi: 10.1109/ACCESS.2019.2919343.

[10] S. A. Priyambada, M. Er, and B. N. Yahya, "Curriculum Assessment of Higher Educational Institution Using Aggregate Profile Clustering," *Procedia Comput. Sci.*, vol. 124, no. 00, pp. 264–273, 2017, doi: 10.1016/j.procs.2017.12.155.

[11] A. A. Mubarak, H. Cao, and I. M. Hezam, "Deep analytic model for student dropout prediction in massive open online courses," *Comput. Electr. Eng.*, vol. 93, no. September 2020, p. 107271, 2021, doi: 10.1016/j.compeleceng.2021.107271.

[12] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," *Mater. Today Proc.*, no. xxxx, 2021, doi: 10.1016/j.matpr.2021.07.382.

[13] X. Lu, Y. Zhu, Y. Xu, and J. Yu, "Learning from multiple dynamic graphs of student and course interactions for student grade predictions," *Neurocomputing*, vol. 431, pp. 23–33, 2021, doi: 10.1016/j.neucom.2020.12.023.

[14] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," *Comput. Electr. Eng.*, vol. 89, no. April 2020, p. 106903, 2021, doi: 10.1016/j.compeleceng.2020.106903.

[15] J. Kuzilek, Z. Zdrahal, and V. Fuglik, "Student success prediction using student exam behaviour," *Futur. Gener. Comput. Syst.*, vol. 125, pp. 661–671, 2021, doi: 10.1016/j.future.2021.07.009.