Profile-Based Cluster Evolution Analysis: Identification of Migration Patterns for Understanding Student Learning Behavior

SATRIO ADI PRIYAMBADA1, (Member, IEEE), MAHENDRAWATHI ER2, BERNARDO NUGROHO YAHYA3, AND TSUYOSHI USAGAWA1, (Member, IEEE)

1Department of Computer Science and Electrical Engineering, Kumamoto University, Kumamoto 860-8555, Japan
2Information Systems Department, Institut Teknologi Sepuluh Nopember, Surabaya 60001, Indonesia
3Industrial and Management Engineering, Hankuk University of Foreign Studies, Yongin, Gyeonggi-do 17035, South Korea

Corresponding authors: Mahendrawathi ER (mahendra_w@is.its.ac.id) and Bernardo Nugroho Yahya (bernardo@hufs.ac.kr)

This work was supported in part by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (JSPS KAKENHI) under Grant 19H1225100.

ABSTRACT: Educational process mining is one of the research domains that utilizes students’ learning behavior to match students’ actual courses taken and the designed curriculum. While most works attempt to deal with the case perspective (i.e., traces of the cases), the temporal case perspective has not been discussed. The temporal case perspective aims to understand the temporal patterns of cases (e.g., students’ learning behavior in a semester). This study proposes modified cluster evolution analysis, called profile-based cluster evolution analysis, for students’ learning behavior based on profiles. The results show three salient features: (1) cluster generation; (2) within-cluster generation; and (3) time-based between-cluster generation. The cluster evolution phase modifies the existing cluster evolution analysis with a dynamic profiler. The model was tested on actual educational data of the Information System Department in Indonesia. The results showed the learning behavior of students who graduated on time, the learning behavior of students who graduated late, and the learning behavior of students who dropped out. Students changed their learning behavior by observing the migration of students from cluster to cluster for each semester. Furthermore, there were distinct learning behavior migration patterns for each category of students based on their performance. The migration pattern can suggest to academic stakeholders to understand about students who are likely to drop out, graduate on time or graduate late. These results can be used as recommendations to academic stakeholders for curriculum assessment and development and dropout prevention.

INDEX TERMS: Educational process mining, cluster evolution analysis, process discovery, trace clustering.

I. INTRODUCTION

Educational process mining (EPM) is a research area that focuses on educational data modeling and analysis to represent students’ learning behavior and match actual students’ learning behavior with curriculum guidelines [1]. By examining students’ learning behavior patterns, we can assess the curriculum and evaluate students’ performance during the study period. Some works on EPM have been applied in academic institutions to predict students’ dropout [2], to recommend a correct path to students [3], to identify, examine and facilitate the evaluation of learning processes [4]–[6], and to enhance the curriculum design [3], [7], [8]. While most existing works emphasize the process perspective and the activity perspective, no previous works attempt to apply the temporal case perspective (i.e., students’ profiles) in accordance with trace (i.e., curriculum) alignment.

In the educational domain, traces can be regarded as course sequences (i.e., curricula) taken by students. Concerning curriculum alignment, it is necessary to consider two types of
curriculum types: noncentralized and centralized curriculum. A noncentralized curriculum is a guideline for students to independently decide on courses without contemplating strict boundaries, such as course prerequisites. On the other hand, a centralized curriculum provides a prerequisite guideline (e.g., Calculus I should be taken before Calculus II) for students to meet academic expectations based on stakeholders (i.e., government and/or policymakers). Many works on EPM consider the noncentralized curriculum, while works on EPM with a centralized curriculum face some challenges regarding the behavior and performance aspects. Concerning the behavior aspect, for example, a student can graduate before the required time if he/she can manage to take courses ahead of schedule. Another challenge involves the performance requirement according to which a student needs to retake courses in the later semesters, which may cause lateness in graduation. Most of the time, the course managers must manually manage courses to be opened each semester, taking into account both students’ behaviors and performance while attempting to achieve the target educational performance, such as a low dropout rate and a high percentage of students graduating on time.

Work in EPM has utilized process mining to identify a process model with limited analysis in profiling students’ behavior. However, the process models of the EPM may be too large and complex for analysis by stakeholders. To analyze various traces (e.g., students’ behaviors), there is an approach to group cases (i.e., instances) based on specific profiles called trace clustering. Originally, trace clustering grouped cases as a full trace without considering partial traces (i.e., time-based trace or temporal case). The result of trace clustering produces homogeneous sets of traces from a full event log and constructs simpler process models [9]–[12]. Some works have used clustering to group students based on specific indicators (i.e., performance and activity) to obtain more specific and accurate models of students’ behavior [13], [14]. In the domain of EPM, the event log can be partially cut in accordance with the timestamp (i.e., period of study or semester). While work has been conducted on trace clustering in the domain of EPM [15], no studies have shown how temporal cases (i.e., students’ profiles) changed over time. Most EPM emphasizes the discovered process model, i.e., the model representing students’ behavior during the study period. However, the evolution of behavior and performance over time remains unexplored. In other words, there is a need to explore and analyze the behavior and performance of students for each semester (called a temporal case) to help stakeholders maintain education quality.

Some previous works have analyzed the evolution of time series data. Shao et al. proposed a new synchronization-based clustering approach for evolving data streams called SyncTree. SyncTree investigates the cluster structure of the data stream between any two timestamps in the past and provides a synchronized microcluster to analyze the cluster evolution [16]. Disegna et al. proposed a method to analyze cluster evolution using repeated cross-sectional ordinal data on tourism market segmentation that helps destination managers and municipalities describe and verify the efficacy of policy and strategies adopted over years without the need to rely on longitudinal surveys, which are often difficult to conduct [17]. Ramon-Gonen and Gelbard proposed the CEA model, which addresses three phenomena likely to occur over time: (1) changes in the number of clusters; (2) changes in cluster characteristics; and (3) between-cluster migration of objects. Among the existing approaches, CEA by Ramon-Gonen is the appropriate method to identify repetition at various points in time and detect patterns that predict prospective loss of value as well as patterns that indicate the stability and preservation of value over time [18]. However, the existing CEA does not consider any mixed profiles during the cluster development stages. In this study, mixed profiles are the alignment course sequence and the respective grade point average (GPA) per semester. The two profiles have different distributions and characteristics at a certain time. Understanding students’ learning behavior in a timely manner requires analysis at a certain point in time (i.e., the semester) to understand the possible actions in the next timestamp (i.e., the next semester).

This study attempts to modify the existing CEA and develop a new method called profile-based cluster evolution analysis (P-CEA). P-CEA considers extracted profiles at a certain time and uses the profiles to understand subsequent learning behavior. In other words, the original profile indicates students’ learning behavior regarding the course sequence, while the additional profile is extracted from the GPA in the respective semester. The P-CEA is expected to determine (1) the amount of data clustered per semester, (2) how cluster characteristics appear per semester, and (3) how cluster migration appears from semester to semester. This study is expected to provide an overview of the learning patterns of students who graduate on time or drop out.

II. RELATED WORK
This section addresses work on EPM and CEA in the domain of educational sectors, such as sequential-based curriculum assessment, trace-based clustering that clusters observation data based on the similarity of the sequences, and relevant works on CEA in various domains.

A. EDUCATIONAL PROCESS MINING
Process mining aims to gain knowledge from event log transactional data. The main tasks of process mining are threefold: 1) discovery, 2) conformance, and 3) enhancement. The event log data were analyzed to discover a process model that best describes the event data. Many works have developed process discovery algorithms to address representational biases and specific problems in particular domains. Conformance checking replays the event log on top of the a priori model to quantify the fitness of the event log for the process model as well as to detect and quantify commonalities or deviations. Enhancement takes the event log and the a priori model to generate a new (enhanced) model. The last task and multiple
perspectives from process data (e.g., organization, control-flow, performance, and resource perspective) can create a simulation model.

In the field of EPM, sequence matching alignment is a curriculum assessment technique [19]. The proposed technique aims to assess the conformity between student learning behavior and curriculum design and groups them into several categories, i.e., match, before, after, repetitive and elective. This research shows that only students’ overall behavior conforms to the curriculum. Cairns et al. proposed methods to understand students’ learning behavior. The methods aimed to discover, analyze, and provide a visual representation of the educational process. The methods used the dotted chart, process modeling, social network, and clustering plug-in on ProM. The results show students’ learning behavior, but they cannot show the conformity of students’ learning behavior with curriculum guidelines. The results can also show the correlations between students’ learning behavior and students’ performance [20].

B. TRACE-BASED CLUSTERING & CLUSTER EVOLUTION ANALYSIS

Trace clustering in process mining can be used for handling scatter and large event logs by dividing them into several groups of clusters with similar types (Song et al., 2013). Trace clustering is different from other process mining algorithms in discovering process models that are usually spaghetti-like, especially from event logs that have highly flexible environments, which are difficult to analyze [21], [22]. Trace clustering can be used to make the obtained model simpler by identifying homogeneous sets [9].

Trace-based clustering for educational data uses sequence matching alignment values to measure the conformity between students’ learning behavior and curriculum guidelines. This technique helps to identify homogeneous sets of traces within a heterogeneous log and makes it easier to discover groups of students’ learning behavior. Priyambada et al. proposed aggregate profile clustering to discover the characteristics of students’ learning behavior and its relationship with students’ performance [23]. Priyambada et al. also proposed segmented-trace clustering that clusters students’ learning behavior per semester. The results showed that the students could be grouped into various clusters per semester that have different learning behavior and performance characteristics. However, the method cannot show the changes in students’ learning behavior in a timely manner (e.g., from one semester to another semester) (Priyambada et al., 2018).

Ramon-Gonen and Gelbard proposed a model called cluster evolution analysis (CEA) that addresses three phenomena that are likely to occur over time: changes in the number of clusters, changes in cluster characteristics, and between-cluster migration of objects. The pattern of the sequence was obtained by using generalized sequential patterns (GSPs). The GSP algorithm is used to obtain frequent sequences from sequence data [24]. The results showed that the model could identify repeated clusters at various points in time and detect pattern migration [18]. This research modified the existing CEA model to identify changes in the learning behavior of students who graduate on time or drop out. This information can provide insights to academic stakeholders for curriculum assessment and development and dropout prevention.

III. METHODOLOGY

This section discusses the methodology of this study. The modified cluster evolution analysis model for the migration of students’ learning behavior is presented in Figure 1. The flow of the analysis is as follows: Data Preprocessing, Profile Assessment, and Cluster Evolution Analysis. Data preprocessing transforms the raw data into event logs and then maps it into projected event logs. Profile assessment includes assessing the behavior log and performance log. Cluster Evolution Analysis involves Temporal Clustering, Cluster-to-State Labeling, Cluster Trajectory, Pattern Recognition, and Pattern Analysis. Each of the phases will be discussed in the next subsection.

A. DATA PREPROCESSING

Data preparation aims to obtain students’ learning behavior from the institution database and prepare the data. Data preparation consists of data collection and sequence matching alignment.

Students’ learning behavior can be captured from students’ performance recorded in academic transcripts. In other words, all information related to students’ academic results in the program is retrieved.

Definition 1: Event Log (L)

An event log (L) is a multiset of traces. Each trace is mapped into a case and consists of a collection of events, which is defined as follows:

- \( E \) is a set of events, and \( E = CS \times T \times G \), where \( CS \) is a set of subjects (courses), \( T \) is a set of timestamps, and \( G \) is a set of grades. A set of courses (CS), as mentioned above, is a collection of the concatenation of course name and course ID.
- Timestamp \((T)\) is the set of the semester on the study program.

The set of grades \((G)\) designated on the program is illustrated by letters such as \( \{A, AB, B, BC, C, D, E\} \) which can be represented as numbers \((4.0, 3.5, 3.0, 2.5, 2.0, 1.0, 0)\), respectively.

Case \((C)\) is a set of events for each student \((ST)\), where a collection of all cases is an event log \((L)\). In other words,
TABLE 1. Event log of students’ course-taking activity.

| Case (C) | Student (ST) | Time-stamp (T) | Courses (CS) | Grade (G) | Credit (CR) |
|----------|--------------|----------------|--------------|-----------|-------------|
| 1 ST001  |              | 2010-1         | Basic Math (CS01) | A | 3 |
| 2 ST001  |              | 2010-1         | Intro to Information System (CS02) | D | 2 |
| 3 ST001  |              | 2010-2         | Intro to Information System (CS02) | B | 2 |
| 4 ST001  |              | 2010-2         | Engineering English (CS03) | E | 3 |
| 5 ST001  |              | 2011-1         | Operation System (CS04) | B | 3 |
| 6 ST001  |              | 2011-1         | Database Management (CS05) | B | 4 |
| 7 ST001  |              | 2011-2         | Engineering English (CS03) | B | 2 |
| 8 ST002  |              | 2010-1         | Basic Math (CS01) | A | 3 |
| 9 ST002  |              | 2010-1         | Intro to Information System (CS02) | A | 2 |
| 10 ST002 |              | 2010-2         | Engineering English (CS03) | A | 3 |
| 11 ST002 |              | 2010-2         | Operation System (CS04) | B | 3 |
| 12 ST002 |              | 2011-1         | Database Management (CS05) | B | 4 |
| 13 ST002 |              | 2011-1         | Information Security (CS06) | A | 3 |
| 14 ST002 |              | 2011-2         | Development (CS07) | B | 3 |
| 15 ST002 |              | 2011-2         | Data Mining (CS08) | A | 3 |
| 16 ST002 |              | 2011-2         | Web Programming (CS09) | C | 4 |
| 17 ST003 |              | 2010-1         | Basic Mathematics (CS01) | A | 3 |
| 18 ST003 |              | 2010-1         | Intro to Information System (CS02) | C | 2 |
| 19 ST003 |              | 2010-2         | Engineering English (CS03) | B | 3 |
| 20 ST003 |              | 2010-2         | Operation System (CS04) | B | 3 |
| 21 ST003 |              | 2010-2         | Database Management (CS05) | B | 4 |

C represents a student’s study history. Cases always have a trace, denoted as $\sigma \in E^*$. $C = E^*$ is the set of possible event sequences (traces describing a case). An example of an event log can be seen in Table 1.

The coarse granularity of the timestamp (i.e., a unit of time such as semester, quarter, or trimester) cannot represent a sequence compared to a finer granularity of timestamps such as hours, seconds, etc. A data structure with a high granularity timestamp is used to analyze the data.

B. PROFILE ASSESSMENT

Before proceeding with the clustering approach, it is necessary to build the feature. This section discusses feature engineering from collected data into prepared data for cluster evolution. First, the event log is projected into the projected event log. The projected event log is aligned with the curriculum guidelines, and it creates a set of sequence alignment features. This set of sequence alignment features will be the feature for cluster evolution.

First, the event log is manipulated into a projected event log.

### Definition 2: Projected event log (PL)

A projected event log is a multiset of traces where events with the same timestamp are grouped into a set. A projected event ($E_t$) is defined as a collection of events that happened in the same timestamp. Note that the sequence in projected events is unordered. A projection function $f: L \rightarrow PL$ where $f(CS) = \{E_t, t = 1, \ldots, n, E_t \subseteq E\}$.

In this study, the projected events contain events that are lexicographically ordered. Hence, the projected event log from Table 1 can be seen in Table 2.

TABLE 2. Converted of a fragment of projected event log.

| Student | Trace |
|---------|-------|
| ST001   | (CS01, CS02) |
| ST002   | (CS01, CS02, CS03) |
| ST003   | (CS01, CS02, CS03, CS04, CS05) |

To align the students’ data with the curriculum guidelines, we utilized sequence alignment matching [19] and produced sequence alignment features. The sequence alignment features contain mixture profiles, alignment with curriculum guidelines, and the grade point average per semester.

### Definition 3: Sequence alignment feature

For example, ST001 finished two courses (CS01 and CS02) in semester 1. Both of the courses aligned with the curriculum guidelines (refer to Table 3). Hence, the values of $d_1^1$, $d_1^2$, $d_1^3$, and $d_1^4$ are 2, 0, 0, and 0, respectively. In semester 3, ST001 took two courses (CS04 and CS05), while the curriculum guidelines enforced CS05 and CS06. In other words, CS05 is aligned with the curriculum guidelines, and CS04 should be taken before semester 3. Hence, the values of $d_3^1$, $d_3^2$, $d_3^3$, and $d_3^4$ are 1, 0, 1, and 0, respectively. The result of the sequence alignment feature according to the data in Table 2 can be seen in Table 4.

### Definition 4: Sequence alignment features according to Table 2.

$$G_i^p = \frac{1}{cr_i^p} \sum p (cr_i^p \cdot G_i^p)$$ (1)

where $cr_i^p$ is the credit of the p-th course in the i-th semester and $cr_i$ is the sum of the credit of all courses in the i-th semester. $G_i^p$ is the grade of the p-th course in the i-th semester.

For example, ST001 finished two courses (CS01 and CS02) in semester 1. The student got an A (4.0) on CS01 with

VOLUME 9, 2021 101721
3 credits and a D (1.0) on CS02 with 2 credits. Hence the GPA-S is \( \frac{1}{3}(4.0 \times 3 + 1.0 \times 2) = \frac{1}{3}(14.0) = 2.8 \).

The profile assessment for the behavior log and performance log can be seen in Figure 2.

Profile assessment consists of behavior profiles and performance profiles. The behavior profile is the sequence alignment matching the behavior log and curriculum guidelines. The performance profile is the performance measure according to the grade and credit numbers of the student.

**Definition 3:** Mixture alignment feature

The mixture alignment feature is a tuple \( D_{m}^i = < d_{i}^l, g_{i} > \), where \( D_{m}^i \) is the set of student learning behaviors and the grade point average of the \( m \)-th student in the \( i \)-th semester. \( d_{i}^l \) is a collection of alignment aggregations in the \( i \)-th semester with 4 categories, \( l = 1, 2, 3, 4 \), named match, after, before, and repetition, respectively. \( g_{i} \) is the grade point average in the \( i \)-th semester. Supposed that ST001 has behavior profile \( d_{2}^2 = 2, 0, 0, 0 \) and performance profile \( g_{2}^2 = 2.8 \), thus student’s mixture profile is \( D_{ST001}^2 = 2, 0, 0, 0, 2.8 \).

**C. PROFILE-BASED CLUSTER EVOLUTION ANALYSIS**

The cluster evolution analysis (CEA) model addresses three phenomena likely to occur over time: (1) changes in the number of clusters; (2) changes in cluster characteristics; and (3) between-cluster migrations of objects. The results showed that the model was capable of identifying repeated clusters at various points in time and detecting patterns that predict prospective loss of value. The CEA model used in this research addresses three phenomena as well, identifying repeated clusters at each semester and detecting patterns that lead to the dropout of students.

The flows of the P-CEA model for the migration of students’ learning behavior are as follows: change in the number of clusters, changes in cluster characteristics, and between-cluster migration of objects.

**1) TEMPORAL CLUSTERING**

The period-based cluster adopts the general clustering method. Clustering is a method to group objects in such a way that similar objects are placed in a group and different objects are placed in another group. In this research, student cluster analysis was used to determine the composition of student learning profiles. One of the popular approaches in clustering is k-Means. K-means clustering is a method that results in a \( k \) cluster, where \( k \) is defined by the user. The group is selected by finding the smallest sum of squared error (SSE) compared to the mean.

The K-means algorithm is used to cluster data per semester, and the elbow method is used to determine the number of clusters. K-means clustering is a method that results in area \( k \) clusters, where \( k \) is defined by the user.

\[
\arg\min_{k} \sum_{i=2}^{k} \sum_{x_{i} \in C_{l}} \|x_{i} - \mu_{j}\|^2
\]

where \( C_{l}^{j} \) is cluster \( j \) in semester \( i \) and \( \mu_{j} \) is the mean point in \( C_{l}^{j} \).

\[
C_{l}^{j} = \{ C_{l}^{j}, l = 2, \ldots |T|, j = 1, \ldots |C_{l}^{j}|, m = 1, \ldots |ST| \}
\]

Various clusters are obtained for each student’s learning behavior data for each semester. There are nine datasets representing the semester. The students’ learning behavior in the 1st semester is not included because in this semester the courses taken by the student are defined by the curriculum. Thus, all students have the same learning behavior, which will result in one cluster in the 1st semester.

**2) CLUSTER-TO-STATE LABELING**

Cluster similarity identification based on cluster characteristics aims to group the cluster into cluster types (states). This step enables us to identify the changes in the cluster characteristics. However, this method does not work properly in data with bounded functions. The number of courses and the course itself is specific to the curriculum guidelines. So the label could be different for each educational organization.

The state is assigned to the categorical label. There are 2 categories: behavior label by using percentage of average matching courses from average total courses of the state, and performance label by using average semester GPA. The behavior label is denoted as \( BL = \{x|x \in N\} \), where \( BL \) is the set of possible behavior and \( x \) is the behavior label. The performance label is denoted as \( FL = \{y|y \in N\} \), where \( FL \) is the set of possible performance labels and \( y \) is the performance label.

\[
\hat{C}_{l}^{j} = (f_{B}, f_{F})
\]

where

\[
f_{B} : C_{l}^{j} \rightarrow BL
\]

\[
f_{F} : C_{l}^{j} \rightarrow FL
\]
dropped out are marked “DO”. Students who graduated are marked “GR”. The cluster trajectory denoted as follows:

\[ CT = \{\hat{CL}_{m,j} | i = 2, \ldots |T|, \]
\[ j = 1, \ldots |\hat{CL}^i|, m = 1, \ldots |ST| \} \]  

(5)

where \( CT \) is Cluster Trajectory, \( \hat{CL}_{m,j} \) is Cluster \( j \)-th in semester \( i \)-th for student \( m \)-th and \( |\hat{CL}^i| \) is the number of clusters in \( i \)-th semester.

Supposed that ST005-ST009 had certain clusters for each semester, so an example of a cluster trajectory can be seen in Table 5. Students ST005, ST007, and ST009 graduated in the 9th semester, and student ST006 graduated in the 8th semester; however, student ST008 dropped out in the 5th semester.

TABLE 5. Example of cluster trajectory.

| Student | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|------|
| ST005  | C1  | C2  | C3  | C5  | C1  | C5  | C1  | GR  |
|         | (H2) | (H3) | (H2) | (M2) | (M2) | (M2) | (M4) | (L1) |
| ST006  | C1  | C1  | C2  | C3  | C1  | C1  | GR  | GR  |
|         | (H2) | (H2) | (M2) | (M2) | (M2) | (M2) | (M1) |
| ST007  | C1  | C2  | C3  | C4  | C1  | C4  | GR  |
|         | (H2) | (H3) | (H2) | (H2) | (H2) | (H4) | (L1) |
| ST008  | C3  | C3  | C4  | DO  | DO  | DO  | DO  | DO  |
|         | (H4) | (M4) | (L4) |     |     |     |     |     |
| ST009  | C1  | C1  | C1  | C1  | C1  | C1  | GR  |
|         | (H2) | (H2) | (H2) | (M2) | (M2) | (M4) | (L1) |

In the former section, we described the process used to identify prominent patterns. A migration graph was designed, providing all the between-cluster student migrations over the entire 9-semester period. Bars represent a period (2nd semester, 3rd semester...), the size of each bar section represents the size of the cluster, and colors represent cluster type (state). The arrows represent the movement (migration) of students. The number near the arrow represents the number of students who migrated.

To analyze the migration of students, we used sequential pattern mining using the generalized sequence pattern (GSP) algorithm to identify prominent patterns between the cluster migration of students. The GSP algorithm determines the frequent pattern of the sequences. The support value for each category denoted as follows:

\[ \text{Support} = \frac{|CT|}{|ST|} \]  

(6)

where

\[ |CT| = |\hat{CL}_{m,j} = \hat{CL}_{n,j} | i, j, m \neq n | m, n = 1, \ldots, |ST| \]  

The length of the frequent sequence can be changed depending on the minimum support value, which is the value that represents the percentage of the same pattern from the overall data. The minimum support of this study tends to be low because the sequence of the data is quite long (9 sequences that represent each semester).

The pseudocode of P-CEA can be seen in Table 6.

IV. RESULTS AND DISCUSSION

A. DATA INTRODUCTION

This section addresses the results of the analysis of a dataset from Information Systems students’ academic data from the 2nd semester to the 10th semester. The first semester data were excluded since all students took the same course in accordance with the curriculum guidelines. The department regularly updates the curriculum every five years. It should be noted that the curriculum guidelines for this study and the batch of students were not changed until the students graduated.

The curriculum guideline is designed for eight (8) consecutive semesters, which is the standard length of study for bachelor’s degrees. During the eight (8) semesters, there are two phases, i.e., the preparation phase (1st semester to 4th semester) and the bachelor’s phase (5th semester to 8th semester). The preparation phase comprises basic courses that are offered to the students from the 1st semester to the 4th semester. Students need to pass all basic courses to continue their study in the 5th semester; otherwise, they drop out. If a student can achieve good performance (i.e., semester GPA ≥ 3.0), he/she can take courses with a total of up to 24 credits. On the other hand, students with poor performance (i.e., semester GPA < 2.0) can take only up to 16 credits. Students who achieve a semester GPA between 2.00 and 3.00 can take up to 20 credits.

B. RESULT

1) TEMPORAL CLUSTERING

By using the k-means algorithm to cluster per semester and the elbow method to determine the number of clusters,
The number of clusters for each semester can be seen in Table 7.

The number of clusters of each semester represents how complex the course-taking pattern is. The number of clusters in the early semester was not high because most of the courses are defined as a set of courses that need to be taken simultaneously. The number of clusters increased in the 5th – 8th semesters because there were more courses, and students freely took courses based on their performance. For example, students with low performance are allowed to take fewer credits and often must retake courses. On the other hand, students with high performance have higher maximum credits and tend to take more semester courses. The last two semesters have the smallest number of clusters because there is no course left in the curriculum guideline and the number of students is small because most of them have graduated.

2) CLUSTER-TO-STATE LABELING
For this case study, we defined three labels (|x| = 3) for behavior label based on its level of conformity to the curriculum guideline that can be seen from percentage of matching courses from total courses. For the performance label, the GPA-S is only bounded to its semester and scaled on a 0-4 value, and we defined the label with four labels (|y| = 4) based on the GPA-S value. The label of each category can be seen in Tables 8 and 9.

The results of state assignments for each cluster using categorical labels are shown in Table 10.

3) CLUSTER TRAJECTORY AND PATTERN RECOGNITION
To obtain a detailed migration pattern, this section shows the migration graph, which is divided into three categories: students who graduated on time, students who graduated late, and students who dropped out.

In Figure 3, which displays the migration graph of students who graduated on time, we find that the majority of students who graduated on time remain in high GPA-S clusters (green and blue) and have high conformity with the curriculum guidelines (Label H). No students who graduated on time migrated or remained in the low GPA-S (red) cluster. From the thickness of the line in the migration graph, we
can see that students who graduated on time had a prominent pattern in the 2nd – 5th semesters and then scattered. The $L$ label, which represents low conformity ($L$) to the curriculum guideline, only appears in the late semesters (7th and 8th semesters).

The GSP algorithm yields sequence patterns from the sequential migration data of students from cluster to cluster for nine semesters. We determined the minimum support parameter to be 0.1 because it yields full sequences (9 sequences). We separated the sequential migration data into three datasets: students who graduated on time, students who graduated late, and students who dropped out. We determined different minimum support parameters for the sequence of students who dropped out to show all the sequences. The results of the sequences for each student category can be seen in Tables 11-13.

| Table 11. Generalized sequence pattern of students who graduated on time.
| Seq (n ST) | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Seq1 (10) | H2  | H2  | H1  | H2  | M2  | M2  | M1  | GR  | GR  |
| Seq2 (7)  | H2  | H2  | M1  | H1  | M2  | L3  | M1  | GR  | GR  |

| Table 12. Generalized sequence pattern of students who graduated late.
| Seq (n ST) | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Seq1 (5)  | H2  | H2  | H2  | M2  | H4  | L1  | GR  |     |     |
| Seq2 (4)  | H2  | H3  | H2  | H2  | M2  | M4  | L1  | GR  |     |
| Seq3 (4)  | H2  | H3  | H2  | H3  | M3  | H4  | L1  | GR  |     |

The sequence patterns show that there are general patterns that differ between students who graduated on time, graduated late, and dropped out. The first and only $L$ (low match) and $3$ (low GPA) occur in the 7th semester for the majority of students who graduated on time. Most of the labels of each cluster for the entire sequence of students who graduated on time consist of $H$ and $M$ as their learning behavior level of conformity and 1 and 2 as their level of semester GPA ranges.

Students who graduated late have three general sequences. Two of them (Seq1 and Seq2) have $H3$ as their cluster label in the 3rd semester, which means that they have high conformity.
with curriculum guidelines but a low average semester GPA. After that, they have stable GPA-S with 2 (green) as their performance label and H and M as their learning behavior label. In the 8th semester, all prominent sequences have 4 (red) as their performance labels due to failure to complete Final Project Course which has 6 credits. For the 9th semester, all prominent sequences have L1 as their profile labels, which means Low on behavior level of conformity (curriculum guideline only provides courses until 8th semester) and good performance. From this pattern, we can conclude that the course in the 8th semester is the biggest problem for the student who graduated late followed by the course in the 3rd semester and 6th semester.

Students who dropped out had 7 different sequences. Five of them had a low average semester GPA, and almost all students dropped out after they had a low semester GPA, except one student (Seq7), indicating that they were unable to catch up with other students. Seq2 and Seq6 only make it until the 2nd semester, Seq3 until the 3rd semester, and Seq4 until the 4th semester while Seq1 and Seq5 still trying until the 5th semester. Seq5 still has a fluctuation of their performance (up in even semester, down in odd semester), on the other hand, Seq1 remains on the 4 (red) performance label, and the behavior label declining due to the increasing number of courses that need to be taken. Only Seq7 did not have red labeled performance but still dropped out in the 3rd semester.

### Table 13. Generalized sequence pattern of students who dropped out.

| Seq | (0 S1) 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
|-----|-----------|-----|-----|-----|-----|-----|-----|-----|-----|
| Seq1 | H4       | M4  | L4  | L4  | DO  | DO  | DO  | DO  | DO  |
| Seq2 | H4       | DO  | DO  | DO  | DO  | DO  | DO  | DO  | DO  |
| Seq3 | H4       | M4  | L4  | L4  | DO  | DO  | DO  | DO  | DO  |
| Seq4 | H4       | M4  | L4  | L4  | DO  | DO  | DO  | DO  | DO  |
| Seq5 | H2       | M4  | H2  | L4  | DO  | DO  | DO  | DO  | DO  |
| Seq6 | L4       | DO  | DO  | DO  | DO  | DO  | DO  | DO  | DO  |
| Seq7 | H2       | H3  | DO  | DO  | DO  | DO  | DO  | DO  | DO  |

### C. Discussion

The temporal case is an interesting perspective in educational process mining because most of the research in this domain focuses on the case perspective. The temporal case perspective can show the change of a pattern in a period of time, especially in a centralized curriculum. A centralized curriculum is a guideline that can be helpful for students to achieve academic expectations by giving a prerequisite course chain. This type of curriculum can also be a problem when students fail a course. Usually, when students fail a course, they need to wait one year to retake the course because of the odd/even semester system. When students fail a prerequisite course, there will be a chain reaction to other courses that can only be taken after they pass the prerequisite course. This situation may prolong the students’ study period.

The main result of this paper is the migration of students from cluster to cluster each semester, which represents how students change their behavior and performance over time. The P-CEA can be a tool to monitor the performance of students by updating the data at the end of the semester. By the end of the semester, department stakeholders should know the position of students’ performance and how they changed their behavior and performance from the previous semester. The cluster in the migration graph is sorted from the highest curriculum guideline conformity and performance to the lowest. From the results, we can see that most of the students who graduated on time tended to migrate to the upper area of the graph; on the other hand, all the students who dropped out tended to migrate to the lower area. The students who graduated late were scattered in all areas of the graph, some students migrated to the upper part of the graph, some others migrated to the lower part. Department stakeholders can provide counseling to students whose performance is declining or who are migrating to the lower area of the graph. For example, the students who migrated in the cluster with low conformity with curriculum guideline (L) and/or low GPA-S4 (red).

Another area that can be analyzed is the occurrence of the cluster. The appearance of a cluster with a low level of conformity in the early semester is because: low-performance students are unable to take the course from the previous semester or retake the course because they failed it previously. Another reason is that students with high performance have already taken the course and need to take the course from the next semester to optimize their maximum credit. Department stakeholders can consider action to help students, such as opening courses that students needed by looking at the occurrence of clusters especially cluster with low GPA-S. By opening a course, students can take/retake that course immediately and should not have to wait until the next year. This type of action will help students who will potentially graduate late.

The difference between P-CEA and CEA is in the data and cluster-to-state labeling process. P-CEA uses educational data of course-taking activity and transforms it to become profile-based data. CEA uses financial data that has different characteristics, such as data distribution and range of values. The educational data used in the P-CEA also has its own characteristic based on their time/semester. The data from earlier semester could be different from last semester. This is because students in the higher semester will have fewer course options as they have taken a lot of courses in the previous semesters.

This methodology works well as can be seen from the migration graph, the differences between migration patterns for each category can be seen clearly. The process in P-CEA is modified process from CEA, so the educational data can be processed, especially in the cluster-to-state labeling process. In the CEA, there is a process that mix the value of the...
This study proposed a profile-based cluster evolution analysis (P-CEA) for dynamic data that occur from time to time. The proposed method was tested on educational data in which the profiles were related to the course sequence alignment and the grade point average of students per semester. The result of the P-CEA provides students’ learning behavior patterns for stakeholders to recognize students who will graduate on time or drop out.

Detailed migration pattern detection shows three migration graphs that indicate the impact of students’ migration per semester on their performance indicated by the study period (students who graduated on time, students who graduated late, and students who dropped out). The migration graph for students who graduated on time shows that students migrated to a better state or remained in the same state with match as their dominant learning behavior and had a good average semester GPA. Students who dropped out remained in the same state or declined in terms of conformity with curriculum guidelines and had a low average semester GPA.

By knowing students’ migration per semester, department management and stakeholders can determine the actions to be taken for students who show declining performance or students who remain in the state with repetitive or later as their dominant learning behavior to prevent students from dropping out by giving the warnings and counseling. Another action that can be taken by stakeholders is to open courses needed by students based on the cluster in the migration graph indicated by the label of the cluster. A limitation of this study is that it is unable to detect students who drop out for reasons outside the lecture. Some students who drop out have good grades and good learning behaviors. Future studies could include students’ demographic data to give more insight into the pattern, exploring another aspect such as social network analysis of course-taking activity between students or lecturers also could be done.

REFERENCES

[1] N. Trcka and M. Pechenizkiy, “From local patterns to global models: Towards domain driven educational process mining,” in Proc. 9th Int. Conf. Infel. Syst. Design Appl., 2009, pp. 1114–1119, doi: 10.1109/ISDA.2009.159.
[2] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, “Predicting students drop out: A case study,” in Proc. 2nd Int. Conf. Educ. Data Min. (EDM), 2009, pp. 41–50, doi: 10.1037/0893-3200.21.3.344.
[3] R. Wang and O. R. Zaiane, “Discovering process in curriculum data to provide recommendation,” in Proc. 8th Int. Conf. Educ. Data Min., 2015, pp. 580–581. [Online]. Available: http://www.educationaldatamining.org/EDM2015proceedings/poster580-581.pdf.
[4] G. Ben-Zadok, A. Hershkovitz, E. Mintz, and R. Nachmias, “Examining online learning processes based on log files analysis: A case study,” in Proc. 5th Int. Conf. Multimedia ICT Educ., 2009, no. 2, pp. 55–59.
[5] B. V. Barreiros, M. Lama, M. Mucientes, and J. C. Vidal, “SoftLearn: A process mining platform for the discovery of learning paths,” in Proc. IEEE 14th Int. Conf. Adv. Learn. Technol., Jul. 2014, pp. 373–375, doi: 10.1109/ICALT.2014.111.
[6] A. Bolt, M. de Leoni, W. M. P. van der Aalst, and P. Gorissen, “Business process reporting using process mining, analytic workflows and process cubes: A case study in education,” in Data-Driven Process Discovery and Analysis (Lecture Notes in Business Information Processing), vol. 244, Cham, Switzerland: Springer, 2017, pp. 28–53, doi: 10.1007/978-3-319-53435-0_2.
[7] W. Y. Wong and M. Lavrenvic, “Using a risk management approach in aligning curriculum for program quality improvement,” in Proc. 1st. Learn. Anal. Curric. Prog. Qual. Improv., 2016, vol. 1, no. 1, pp. 10–14.
[8] N. Patel, C. Sellman, and D. Lomas, “Mining frequent learning pathways from a large educational dataset,” 2017, arXiv:1705.11125. [Online]. Available: http://arxiv.org/abs/1705.11125
[9] J. Evermann, T. Thaler, and P. Fettek, “Clustering traces using sequence alignment,” Business Process Management Workshops (Lecture Notes in Business Information Processing), vol. 256, Cham, Switzerland: Springer, 2016, pp. 179–190, doi: 10.1007/978-3-319-42887-1_15.
[10] J. De Weerdt, S. K. L. M. van den Broucke, J. Vanthienen, and B. Baensens, “Active cluster mining for improved process discovery,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2708–2720, Dec. 2013, doi: 10.1109/TKDE.2013.64.
[11] Y. Van Oirschot, B. Van Dongen, J. Buijs, and R. Dijkman. (Aug. 2014). Using Trace Clustering for Configurable Process Discovery Explained by Event Log Data. Processmining.org. [Online]. Available: http://www.processmining.org/_media/blogs/pub2014/thesisyvoirschot_final.pdf
[12] J. F. A. Hopmes, J. Buijs, W. M. P. van der Aalst, P. M. Dixit, and J. Buurman, “Discovering deviating cases and process variants using trace clustering,” in Proc. 27th Benelux Conf. Artif. Intel. (BNAIC), Nov. 2015, pp. 5–6.
[13] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” Expert Syst. Appl., vol. 41, no. 4, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.
[14] H. Ariouat, A. H. Cairns, K. Barkaoi, J. Akoka, and N. Khelifa, “A two-step clustering approach for improving educational process model discovery,” in Proc. IEEE 25th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises (WETICE), Jun. 2016, pp. 38–43, doi: 10.1109/WETICE.2016.18.
[15] S. A. Priyambada, M. Er, and B. N. Yahya, “Curriculum assessment of higher educational institution using segmented-trace clustering,” Jurnal Teknik Industri, vol. 20, no. 1, pp. 33–48, Jun. 2018, doi: 10.9744/jti.201.1.33-48.
[16] J. Shao, Y. Tan, L. Gao, Q. Yang, C. Plant, and I. Assent, “Synchronization-based clustering on evolving data stream,” Inf. Sci., vol. 501, pp. 573–587, Oct. 2019, doi: 10.1016/j.ins.2018.09.035.
[17] M. Diuegna, P. D’Urso, and R. Massari, “Analysing cluster evolution using repeated cross-sectional ordinal data.” Tourism Manage., vol. 69, pp. 524–536, Dec. 2018, doi: 10.1016/j.tourman.2018.06.028.
[18] R. Ramon-Gonen and R. Gelbard, “Cluster evolution analysis: Identification and detection of similar clusters and migration patterns,” Expert Syst. Appl., vol. 83, pp. 363–378, Oct. 2017, doi: 10.1016/j.eswa.2017.04.007.
[19] L. Y. Bendatu and B. N. Yahya, “Sequence matching analysis for curriculum development,” Jurnal Teknik Industri, vol. 17, no. 1, pp. 47–52, Jun. 2015, doi: 10.9744/jti.17.1.47-52.
[20] A. H. Cairns, B. Guent, H. Hafidi, C. Joubert, and N. Khelifa, “Towards a distributed computing platform tailored for educational process discovery and analysis,” in Proc. Int. Conf. Protocol Eng. (ICPE) Int. Conf. New Technol. Distrib. Syst. (NTDS), Jul. 2015, pp. 1–8, doi: 10.1109/NOTERE-2015.72934.94.
[21] J. De Weerdt, A. Schupp, A. Vanderloock, and B. Baensens, “Process mining for the multi-faceted analysis of business processes—A case study in a financial services organization,” Comput. Ind., vol. 64, no. 1, pp. 57–67, Jan. 2013, doi: 10.1016/j.compind.2012.09.010.
[22] M. No and J. Liu, “A profile clustering based event logs repairing approach for process mining,” IEEE Access, vol. 7, pp. 17872–17881, 2019, doi: 10.1109/ACCESS.2019.2894905.
[23] S. A. Priyambada, E. R. Mahendrawati, and B. N. Yahya, “Curriculum assessment of higher educational institution using aggregate profile clustering,” Procedia Comput. Sci., vol. 124, no. 1, pp. 264–273, 2017, doi: 10.1016/j.procs.2017.12.155.
[24] A. K. Pujari, Data Mining Techniques. Hyderabad, India: Universities Press, 2001.
SATRIO ADI PRIYAMBADA (Member, IEEE) received the bachelor’s and master’s degrees in information systems from the Institut Teknologi Sepuluh Nopember, Indonesia. He is currently pursuing the Ph.D. degree in computer science with the Graduate School of Science and Technology, Kumamoto University, Japan. He is a member of the Human Interface Cyber Communication Laboratory. His research interests include process mining, machine learning, and educational data mining.

MAHENDRAWATHI ER received the bachelor’s degree in industrial engineering from the Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, and the master’s degree in operations management and manufacturing systems and the Doctoral degree in manufacturing engineering and operations management from Nottingham University, Nottingham, U.K. Since 2006, she has been a Lecturer with the Information Systems Department, Institut Teknologi Sepuluh Nopember. She teaches among others enterprise resource planning, business process management, supply chain management, and e-business. She has published her work in international journals and conferences. Her research interests include business process management, supply chain management, and enterprise systems. She is a member of the Enterprise Systems Research Laboratory.

BERNARDO NUGROHO YAHYA received the B.S. degree in industrial engineering from Petra Christian University, the M.S. degree in information systems from Dongseo University, and the Ph.D. degree in industrial engineering from Pusan National University. He is currently a Full Professor with the Industrial and Management Engineering Department, Hankuk University of Foreign Studies, South Korea. He has been working on various industry business consulting and engineering projects with Korean companies. His current research interests include statistical pattern recognition, machine learning, business process intelligence, and data analytics.

TSUYOSHI USAGAWA (Member, IEEE) received the B.E. degree from the Kyushu Institute of Technology, in 1981, and the M.E. degree from Tohoku University, in 1983. Since 1983, he has been with Kumamoto University, where he has been a Professor with the Graduate School of Science and Technology, since 2004. From 2014 to 2020, he was the Vice Dean of the Graduate School of Science and Technology, Kumamoto University, Japan. He is currently acting as the Trustee Vice President of Kumamoto University. His research interests include acoustic signal processing, perceptual information processing, educational data mining, and e-learning among others. He is a member of ASA, ASI, INCE/EI, IEICE, JSET, JAIS, and ACM. From 2005 to 2007, he was the Vice President of ASJ.

...