# Evolutionary Clustering of Apprentices' Behavior in Online Learning Journals for Vocational Education

Paola Mejia, Mirko Marras, Christian Giang, Alberto Cattaneo and Tanja Käser

## 1 Dimensions of Self-Regulated Learning

### 1.1 Data Pre-processing

The platform tracks ten events per apprentice: log in, create recipe, edit recipe, create experience, edit experience, create reflection, edit reflection, upload image, request feedback, receive feedback. These event are described in Section **??**.

This set of possible actions does not contain a log-out event, which makes it hard to estimate the time apprentices spent on the platform (time-on-task estimation). For instance, if an apprentice takes a long break after writing a recipe and then starts editing the recipe again four hours later, it could seem that the apprentice worked on that recipe for four hours, while in reality the apprentice spent only ten minutes on it. Kovanovi *et al.*[1] refer to this problem as an outlier detection problem where the outliers are the unusually long activities. They note that a typical solution for dealing with these cases is to put a maximum value on the duration of activities (for example, 10-15 minutes or one hour). In our work, we noticed that the time spent on tasks varied between the apprentices and the tasks. For example, some apprentices regularly spent 20 minutes writing the recipe and the reflection whereas other students spent 2 minutes editing the recipes. Thus, we set a maximum per apprentice and per activity. To set this maximum, we used the boxplot definition of outlier (1.5xIQR) [2]. We first calculate the distribution $\phi_u(\Delta t = t_{i+1} - t_i)$ of the time between two consecutive events per activity per apprentice. Next, we define maximum as follows:

$$maximum = Q3_{\phi_u(\Delta t)} + 1.5 \cdot (Q3_{\phi_u(\Delta t)} - Q1_{\phi_u(\Delta t)}), \tag{1}$$

**Event Sequencing**. We extract the events $e_i$ for each user (apprentice) $u$ with $e_i = 1, ..., N_u$, where $N_u$ denotes the total number of events of user $u$. We assume that each event $e_i$ is represented by a tuple $(t_i, a_i, o_i)$, including a timestamp $t_i$, an action $a_i$, and a resource $o_i$. In case of the platform of this study, events can be either recipes ($rc$) or experiences ($ex$), but the notation can be easily extend to other types. As described in Section **??**, both recipes and experiences come attached with metadata: title, description, reflection, ingredients (only for recipes), tags (denoting the associated topic), images, feedback requests, and received feedback. We use superscripts to denote an attribute of a resource. The ingredients of a recipe will for example be referred to as $rc^{ingredients}$ and $o^{reflection}$ will describe the reflection of a resource $o$ (recipe or experience).

**Time frame and time units**. Finally, we introduce the concepts of *time frame* $tf$ and *time unit* $ts$. A time frame denotes the period of time for which we want to identify apprentices' profiles. A time unit denotes the smallest unit of time for which we compute features. The time frame can for example be a year, or semester, and the time unit a biweek, or week. We denote time frames and units as superscripts, e.g., $\mathcal{S}^{tf,ts}$ are the sessions of user $u$ in time step $ts$ of time frame $tf$.

### 1.2 Feature Engineering

Some of the features are represented by a single *scalar* value for each time frame. Other features are represented by a *time series*: they are computed per time unit and therefore consist of $|\mathcal{T}|$ feature values over the time frame, where $|\mathcal{T}|$ denotes the number of time units within a time frame. Moreover, each dimension $\mathcal{F}_i$ is composed by a set of features $\mathcal{F}_i = f_{i,1}, \ldots, f_{i,k}$. For example, using semester as time frame and biweek as time unit, we obtain a sequence of length 13; in which every value in the vector refers to a two-week period. Explaining the notation further, in the previous example, $f_{2,1}^{4,3}$ is a scalar that refers to the 1st feature of the 2nd dimension, in the 3th biweek of the 4th semester. To simplify the notation, from this point on, we assume that all the formulas are for user $u$. Furthermore, all the features except for the regularity features, are time series and are computed for each time unit $ts$ of a time frame $tf$. The regularity feature are scalars and are only computed once every time frame $tf$.

**Effort Dimension** $\mathcal{F}_1$. We then consider two types of absolute effort measurements; both are time series features and hence computed per time unit $ts$ of a time frame $tf$. The first feature $f_{1,1}$ denoting the *number of minutes spent on the platform* is defined as follows:

$$f_{1,1}^{tf,ts} = \sum_{s \in \mathcal{S}^{tf,ts}} (t_{\text{s, last}} - t_{\text{s, first}})_{\text{minutes}}, \tag{2}$$

where $S^{tf,ts}$ is the set of sessions within the time unit $ts$ of a time frame $tf$, and $t_{s,\text{last}}$ and $t_{s,\text{first}}$ are the timestamps of the last and first event of session $s$. The second feature $f_{1,2}$, describing the *number of writing events* is computed as:

$$f_{1,2}^{tf,ts} = \sum_{a \in \mathcal{A}^{tf,ts}} 1(a \in \{create\_*, edit\_*\}), \tag{3}$$

where $\mathcal{A}^{tf,ts}$ are the set of actions in time step $ts$ of a time frame $tf$, and $1(a \in \{create\_*, edit\_*\})$ is 1 if $a$ is a create or edit action of an experience or recipe, and 0 otherwise.

**Quality Dimension** $\mathcal{F}_2$. We consider four different features as a proxy of the quality of the content. Feature $f_{2,1}$ denotes the *the average length of reflections* and is defined as follows:

$$f_{2,1}^{tf,ts} = \frac{1}{|\mathcal{O}^{tf,ts}|} \sum_{o \in \mathcal{O}^{tf,ts}} |o^{reflection}|, \tag{4}$$

where $\mathcal{O}^{tf,ts}$ denotes the set of resources that the apprentice manipulated in time unit $ts$ of a time frame $tf$ and $|o^{reflection}|$ is the number of characters of the reflection text of resource $o$. The second feature $f_{2,2}$ computes the *the ratio of recipes with at least one ingredient* as follows:

$$f_{2,2}^{tf,ts} = \frac{1}{|\mathcal{R}_u^{tf,ts}|} \sum_{o \in \mathcal{R}_u^{tf,ts}} 1(|rc^{ingredients}| > 0), \tag{5}$$

where $\mathcal{R}^{tf,ts}$ denotes the set of recipes manipulated in time unit $ts$ of a time frame $tf$ and $1(|rc^{ingredients}| > 0)$ is 1 if the recipe $rc$ has more than one ingredient and 0 otherwise. The third and fourth features $f_{2,3}$ and $f_{2,4}$ are the *ratio of recipes and experiences with at least one image and tag*, respectively:

$$f_{2,3}^{tf,ts} = \frac{1}{|\mathcal{O}^{tf,ts}|} \sum_{o \in \mathcal{O}^{tf,ts}} 1(|o^{images}| > 0), \tag{6}$$

$$f_{2,4}^{tf,ts} = \frac{1}{|\mathcal{O}^{tf,ts}|} \sum_{o \in \mathcal{O}^{tf,ts}} 1(|o^{tags}| > 0), \tag{7}$$

where $\mathcal{O}^{tf,ts}$ are resources that the apprentice manipulated in time unit $ts$ of a time frame $tf$, and $1(|o^{images}| > 0)$ and $1(|o^{tags}| > 0)$ return 1 if the resource $o$ has more than one image and tag respectively, and 0 otherwise.

**Consistency Dimension $\mathcal{F}_3$.**

The first feature $f_{3,1}$ measures the *average session duration*:

$$f_{3,1}^{tf,ts} = \frac{1}{|\mathcal{S}^{tf,ts}|} \sum_{s \in \mathcal{S}^{tf,ts}} (t_{\text{r, last}} - t_{r,\text{first}})_{\text{minutes}}, \tag{8}$$

where $\mathcal{S}^{tf,ts}$ is the set of sessions within time unit $ts$ of a time frame $tf$ and $t_{\text{r, last}}$ and $t_{r,\text{first}}$ are the timestamps of the last and first event in a session $s$. The second feature $f_{3,2}$ denoting *the relative use of the platform* is defined as follows:

$$f_{3,2}^{tf,ts} = \frac{1}{t_{total}^{tf}} \sum_{s \in \mathcal{S}^{tf,ts}} (t_{\text{last}} - t_{\text{first}})_{\text{minutes}}, \tag{9}$$

where $t_{\text{last}}$ and $t_{\text{first}}$ are the timestamps of the last and first event in a session $s$, and $t_{total}^{tf}$ is the total amount of time (in minutes) the apprentice spent on the platform in time frame $tf$. The last feature $f_{3,3}$ of this dimension computes the *relative number of writing events* $f_{3,3}$ as follows:

$$f_{3,3}^{tf,ts} = \frac{1}{|\mathcal{E}^{tf}|} \sum_{a \in \mathcal{A}^{tf,ts}} 1(a \in \{create\_*, edit\_*\}) \tag{10}$$

where $\mathcal{A}^{tf,ts}$ are the actions performed in time unit $ts$ of a time frame $tf$ and $1(a \in \{create\_*, edit\_*\})$ is 1 if $a$ is a create or edit action of a recipe or experience and 0 otherwise. $\mathcal{E}^{tf}$ is the number of writing events during time frame $tf$.

**Help-Seeking Behavior Dimension $\mathcal{F}_4$.**

We describe apprentices' help-seeking behavior by two features, which we again compute per time unit $ts$ within a time frame $tf$ (time series features). Specifically, we define the first feature $f_{4,1}$ as *the relative amount of feedback requests*:

$$f_{4,1}^{tf,ts} = \frac{|\{o \in \mathcal{O}^{tf,ts} |\ o^{\text{request\_feedback}} > 0\}|}{|\mathcal{O}^{tf,ts}|}, \tag{11}$$

where $\mathcal{O}^{tf,ts}$ is the list of resources (with repetition) manipulated in time unit $ts$ of time frame $tf$. The second feature $f_{4,2}$ denotes the *relative amount of answers to feedback requests*:

$$f_{4,2}^{t,s} = \frac{|\{o \in \mathcal{O}^{t,s} |\ |o^{\text{receive\_feedback}}| > 0\}|}{|\{o \in \mathcal{O}^{t,s} |\ o^{\text{request\_feedback}} > 0\}|}, \tag{12}$$

where $\mathcal{O}^{tf,ts}$ is the list of resources (with repetition) manipulated time unit $ts$ of a time frame $tu$, and $|o^{\text{receive\_feedback}}|$ is the number of characters included in the feedback answer of resource $o$ (non-empty if the feedback was provided).

**Regularity Dimension $\mathcal{F}_5$.**

To assess apprentices' regularity, we adapted the regularity features to fit to our time units. These features are all scalars that are computed over the course of the time frame $tf$. The first feature $f_{5,1}$ describes the tendency to work more on *specific days* of the time unit $ts$:

$$f_{5,1}^{tf} = (\log(ts_{days}) - E_W) \max_d W(d), \tag{13}$$

where $ts_{days}$ indicates the number of days $d$ in time unit $ts$, $W$ is a vector of length $ts_{days}$, with $W(d)$ indicating the sum of all actions performed on day $d$ over each time unit $ts$ in time frame $tf$, and $E_W$ is the entropy of the distribution $W$. This feature is maximized, when the distribution $W$ has a large peak, i.e. a large number of actions on a day $d$ of the time unit and a small number on all other days (small entropy $E_W$). It is minimized in case of a uniform distribution ($E_W = log(ts_{days})$). This feature tells if actions are focused around a day of the time unit (e.g. the apprentice is more active on Tuesdays). The second feature $f_{5,2}$ describes the tendency to work *on certain hours of the day*:

$$f_{5,2}^{tf} = \text{FFT}_{1/24}(W), \tag{14}$$

where $FFT$ indicates a Fast Fourier Transformation, 24 indicates the number of hours $h$ in a day, $ts_{days}$ indicates the number of days $d$ in time unit $ts$, and $W$ is a vector with $W(x = hd)$ indicating the number of actions on a specific hour $h$ of the day $d$ over the time frame $tf$. The subscript $1/24$ indicates the frequency resolution used in $FFT$, that is the temporal period where the periodicity of the distribution $W$ is checked (i.e., every 24 hours). This feature is maximized, when the periodicity is high, i.e. the same number of actions performed on a each specific hour $h$ every day in the time frame $tf$. It is minimized in case of a low periodicity, i.e., the hourly pattern highly changes over days in the time frame

$tf$. This feature measures the extent to which the hourly pattern of activities is repeating over days (e.g. the apprentice is active from 8h-10h and 12h-17h on every day). Finally, the third feature $f_{5,3}$ denotes the tendency to work on *on specific hours* ***and*** *days of the time unit*:

$$f_{5,3}^{tf} = \mathrm{FFT}_{1/(24 \cdot ts_{days})}(W), \tag{15}$$

where the subscript $1/(24 \cdot ts_{days})$ indicates that the periodicity is checked across hours 24 and days $ts_{days}$ in a time unit. This feature is maximized when we have the same number of actions on a each hour $h$ of each day $d$ in the time unit $ts$, over time units. It is minimized in case the number of actions per hour and day in a time unit $ts$ changes. This feature tells if the hourly-daily pattern is repeating over time units (e.g., in every time unit, the apprentice is active at 8h-10h on Monday, 12h-17h on Fridays).

# References


[1] V. Kovanović, D. Gašević, S. Dawson, S. Joksimović, R. S. Baker, and M. Hatala, "Penetrating the black box of time-on-task estimation," in *Proc. of the Fifth Int. Conf. on Learning Analytics And Knowledge*, ser. LAK '15. New York, NY, USA: Association for Computing Machinery, 2015, doi: http://dx.doi.org/10.1145/2723576.272362310.1145/2723576.2723623. p. 184–193.

[2] J. W. Tukey *et al.*, *Exploratory data analysis*. Reading, Mass., 1977, vol. 2.