Identification of Class-Representative Learner Personas

Célina Treuillier [1,2] and Anne Boyer[1,2]

[1] *Lorraine University, 34 Cours Léopold, 54000 Nancy, France*
[2] *LORIA, 615 Rue du Jardin Botanique, 54506 Vandœuvre-lès-Nancy, France*

Abstract

The student's interaction with Virtual Learning Environments produces a large amount of data, known as learning traces, which is commonly used by the Learning Analytics (LA) domain to enhance the learning experience. We propose to define personas, that are representative of subsets of students sharing common digital behaviors. The embodiment of the output of LA systems in the form of personas makes it possible to study the representativeness of the dataset with precision and act accordingly, but also to enhance the explicability to pedagogical experts who must manipulate these tools. These personas are defined from learning traces, which are processed to identify homogeneous subsets of learners. The presented methodology also allows to identify some outliers, that exhibit atypical behaviors, and thus makes it possible to represent the whole students, without privileging some of them.

Keywords

Learning Analytics – Learning Systems – Learner Personas – Virtual Learning Environments – Explicability – Corpus representativeness.

1. Introduction

The generalization of digital environments in education leads to the collection of big amounts of educational data, which can either be personal information on learners, academic performances of students, or interaction traces. This data could be processed by Learning Analytics (LA) tools. LA was defined in 2011 as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (1). It allows to understand the digital behaviors of students, to model, explain, or predict them, and then to better understand the use of a smart learning environment (SLE).

The collection and exploitation of educational data lead to ethical questions such as privacy, security, informed consent, or bias (2). Some specific frameworks have been proposed, such as the DELICATE checklist (3) which provides a guide for assessing the proper use of educational data. More recently, some researchers (4) mention a need for a more complete and accurate evaluation of digital learning environments, going beyond the common evaluation which mainly deals with global algorithmic performances. Even if the computation of various measures (precision, recall, RMSE, MAE...) gives clues about the quality of the system (5), more pedagogical aspects are missing. This paper is a contribution to the design of a methodology dealing with the critical issue of the automatic identification of digital learning behaviors from educational data. Of course, knowing these digital learning behaviors leads to a more precise evaluation (performances can be given for each specific digital learning behavior). It may also give information to pedagogical experts on the way learners behave within a specific SLE, and therefore contributes to explicability.

In this context, we propose to characterize learners' online behaviors using learning indicators reflecting the behaviors (interaction, activity, learning) of a specific learner. They are computed from a subset of features available in learning traces and bring significant pedagogical information (6,7). Measuring such indicators makes it possible to differentiate students based on their behaviors, and thus to provide them with more personalized support (8). Indeed, within a single class, not all students have the same needs and advice is not appropriate for all learners, especially in large groups in which students have varied backgrounds, objectives, and skills (9). All the more so since this lack of homogeneity among students is exacerbated in an online learning environment, which increases inequality (9).

Several studies have already attempted to categorize students based on learning traces for different purposes: to identify students who can benefit from the same intervention by the instructor (10), to detect students who are going to drop out or students at risk (11,12), to evaluate performance (13), to provide adapted recommendations (14)... Here, we are more interested in defining online behaviors in order to characterize the dataset in a new way. That is why we propose to define the latter in the form of "personas", corresponding to subsets of students sharing common behaviors. The description of the dataset in the form of personas will first allow us to analyze the representativeness of the corpus: the learning performances will be detailed according to the various subsets of students, and it will be possible to evaluate if some are under-represented, over-represented, or not represented, for example. But these personas will also allow improving the explicability by embodying the outputs of the system in the form of fictitious students to whom pedagogical experts can refer.

The challenge is therefore to be able to define learner personas from the learning traces. The research question is then **(RQ) How to define learner personas based on learning traces and indicators?** To carry out this study, we work with the broadly used Open University Learning Analytics Dataset (OULAD), which is described in the following part. We present our methodology in the third section. The results are described in the fourth part. Finally, we conclude and give some perspectives.

2. Dataset and learning indicators

The OULA Dataset (15) gathers data about 32,593 students involved in distance learning. It is fully anonymized and contains both demographic data, interaction data, as well as the results of the various evaluations. The interaction data mainly focused on the activity on available materials, i.e., the clicks made on specific resources, and are time stamped. Students may have 4 types of outcomes: pass, fail, withdrawn, or distinction. We select the presentation of February 2013 of the STEM module D (duration of 240 days, 14 assessments, 1303 students). As previously explained, the division of students into subsets sharing common digital behavior is based on learning indicators that we characterize from some existing studies. In total, 5 indicators are used: engagement (16), performance (17), regularity (18), responsiveness (18), and curiosity (19). *Table 1* summarizes the description of the indicators.

Table 1
Learning indicators

Indicator	Definition	Features
Performance	Student's outcomes	Scores in the 14 assessments, ranging from 0 to 100.
Reactivity	Responsiveness to course-related events	Delay between the date the assignment is returned and the deadline (in days).
Engagement	Student activity	Number of clicks on selected types of activities + Total number of clicks all activities combined.
Regularity	Behavioral patterns of actions	Number of active days on selected types of activities + Total active days + Mean of the number of clicks per day on the same types of activities and global
Curiosity	Intrinsic motivation	Number of different types of activity consulted + Number of different resources consulted.

When students did not turn in an assignment, did not get a grade for an assessment, or did not make any click, the initial dataset includes null values. We replace missing values with 0 when no clicks were made, or no results were indicated. Alike, when an assignment was not returned, we replace the missing values by 240, corresponding to the duration of the course. As resources are available a few weeks before the course starts, some students have a number of active days superior to the duration of the module, up to 260 days. Our initial dataset D thus included a total of 45 features corresponding to the description of the 5 learning indicators, for 1303 students. We divide D according to the 4 types of

results and obtain 4 independent datasets whose size is summarized in *Table 2*. Each dataset is analyzed and thus undergoes various processing steps, which are detailed in the following section.

Table 2
Dimensions of the four datasets

Dataset	Number of students	Proportion
Pass	456	35,0%
Fail	361	27,7%
Withdrawn	432	33,2%
Distinction	54	4,1%

3. Methodology

To meet our challenges (evaluation of representativeness and enhancement of explicability), we propose to define homogeneous subsets of students adopting similar behaviors from a heterogeneous set. Each student is characterized by his profile, consisting of a sequence of learning traces. Some students present typical behaviors and cannot be associated with any sufficiently large subset. They are therefore considered as 'outliers' and are treated separately.

The initial dataset D is composed of several learners described by their profiles $P_1, P_2, \dots, P_n$. Each profile P_i, associated with a single student, is composed by a sequence of traces $T_{i,j}$ (*j-th trace of student associated with the profile* P_i). The goal is to find homogeneous subsets S_k, i.e., subsets of profiles P_i composed by sequences of traces reflecting similar behaviors. Profiles that are too dissimilar are therefore considered as outliers O_p. If the number of profiles P_i in a subset S_k is lower than a threshold ε, we consider the associated profiles as outliers.

We will use these subsets to describe "personas", which have been defined by Brooks and Greer (20) as “narrative descriptions of typical learners that can be identified through centroids of machine learning classification processes”. In our case, learner personas will be based on student interaction data with the learning environment and are defined from outcomes of the clustering method. It is important to note that our definition of personas differs from the one commonly used in UX design (21). Indeed, here, personas are used after the design phase of the tool to ensure that the latter can respond to all students with the same quality. Thus, the personas we define allow us to describe a digital learning behavior shared by several students likely to benefit from the same advice, to study the representativeness of the corpus, and to enhance the explicability.

The applied methodology is broken down into different parts*:* first, the data undergoes a pre-processing phase during which we handled the null values (NAs) and standardized the data. Data standardization is a common process applied in Machine Learning, allowing to resize numerical variables to make them comparable on a common scale. After this pre-processing phase, we detect outliers: this allows splitting the initial dataset into inliers dataset and outliers dataset. Due to their atypical behavior, the outliers are examined independently, and the inliers are divided into subsets using an unsupervised clustering algorithm. Finally, the characteristics of each homogeneous group, i.e., the behaviors adopted, allow the definition of personas, which are descriptions of typical students to whom the system must be able to respond, and always with the same quality.

4. Experimentation

4.1. Description

The whole implementation was performed using the Scikit Learn library for Python (15). For the standardization phase, after studying and comparing the different existing scalers, we selected the

RobustScaler scaler proposed by ScikitLearn which is particularly adapted for datasets including outliers. We then applied the IsolationForest algorithm to isolate atypical data, with contamination set to 0,01. Finally, we processed the K-means algorithm, which is adapted for LA datasets (22), for the clustering phase. The centers of resulting clusters allow us to define the personas and analyze them. The quality of the partition is evaluated using the Davies-Bouldin criterion (13), and Silhouette analysis (14). All the steps were applied independently on our four datasets (Pass, Fail, Withdrawn, Distinction).

4.2. Results

First, the IsolationForest algorithm allows to identify inliers and outliers, and therefore separate them into independent datasets. The number of outliers and inliers for each dataset is given in *Table 3*.

Table 3
Number of inliers and outliers

Dataset	Inliers	Outliers
Pass	451	5
Fail	357	4
Withdrawn	427	5
Distinction	53	1

Inliers were then processed with the K-means algorithm for different values of K (2 to 10,12,15) and performance measures were computed to choose the optimal number of clusters *(Table 4).*

Table 4
Dimensions of the four datasets

Dataset	Optimal value of K	Davies-Bouldin Index	Silhouette Index
Pass	10	0,70	0,78
Fail	8	0,16	0,91
Withdrawn	4	0,82	0,83
Distinction	6	0,05	0,88

For each dataset, clusters sizes, i.e., the number of students sharing similar behaviors within the same subset, differ greatly. Overall, in each dataset, there is a larger subset representing the major proportion of learners, and some smaller subsets, sometimes representing only one student. The larger subset was defined as the prime persona: it is representative of the majority of students in the studied dataset. Smaller clusters (size > ε) were defined as under-represented personas. Please note that these personas, even if they represent fewer learners, need to be evaluated and treated with the same quality as a prime persona. Finally, as explained, the students composing clusters of size smaller than $\varepsilon = 10$ are considered as outliers. These last exhibit unique behaviors and need to be treated separately because they must require adapted support, as those identified with the IsolationForest algorithm.

In this paper, due to lack of space, we cannot describe all the personas, but we detail the most interesting and representative ones and give relevant values corresponding to the clusters' centers of the described persona. Firstly, for successful students, the primary persona represents 69% of the dataset (312 learners). These students are very active (2240 clicks), especially on the forums (522 clicks). They are also regular since they are active for more than 130 days over the total duration of the module. The resources consulted are numerous (167). This active, regular, and curious behavior allows them to obtain good results throughout the module. Other students, less represented, are less active with half the number of clicks (1113) and far fewer active days (77). These students, less active and less regular, do not turn in all the assignments but their correct results nevertheless allow them to validate the

module. Finally, the outliers include students with phrenic activity (19196 clicks) spread over 259 active days during which 439 different resources are consulted. We can easily understand why this type of student is considered as outliers given the adopted behaviors.

If we now consider the students who failed (53% of the dataset, 190 students), the majority of them are not very active (620 clicks), whatever the type of activity considered. This low activity is associated with a reduced number of resources consulted (73) and less active days (43). These students who are not very active, irregular, and not very curious about the course, obtain low results that do not allow them to succeed, especially since they are not very reactive and do not return all the assignments. However, other under-represented students were more active (1871 clicks), more regular (110 active days), more curious (145 resources consulted), and turned in all the assignments on time but obtained low scores and therefore performed poorly. The work provided does not seem to allow this subset of students to succeed. Interestingly, we observe that some outliers have a sustained activity with a large number of clicks, many days of activity, and a wide variety of resources consulted, but obtained scores are too low to pass the module.

Next, we observe that the majority of students who dropped out (76 % of the dataset, 326 students) have very low activity (351 clicks), are very irregular (22 active days), and access very few resources (45). This behavior causes poor results from the beginning of the course. These students dropped out quickly, and do not turn in any more assignments. Other underrepresented subgroups are more active (number of clicks > 1000), more regular, and more curious, but give up more or less quickly. One of the outliers of this dataset shows an exemplary behavior at the beginning of the course with high activity (4267 clicks), a high regularity (178 active days), and curiosity (188 consulted resources), but gives up for the last assignment, which is not handed in.

Finally, the majority of students earning a distinction (87%, 46 students) are very active (2577 clicks) and regular (146 active days) in the course. In particular, they show high activity on the forums (627 clicks). This behavior allows them to obtain excellent results. For this dataset, we do not observe any under-represented personas. Students not belonging to the main subset are outliers. The most different of them is an outlier showing a very increased activity (17957 clicks) throughout the entire course (260 days of activity) and a high curiosity (361 resources consulted). This student also seems to be very active on the forums since he makes almost 7050 clicks in it. All of his assignments are handed in on time and his results are excellent.

The described personas are interesting since they are diversified and allow to clearly differentiate the students according to their online behaviors. Besides, the personas of each dataset are very representative of the associated final result. Thus, the subsets of students identified as a result of our methodology are representative of a variety of digital behaviors, and therefore do not focus on describing the most common ones. In this way, the representativeness analysis of the corpus can be improved, ensuring that students engaging in underrepresented behaviors are identified and treated with the same quality as other students. Finally, the association of each persona with various learner indicators makes it possible to embody the results of LA algorithms in a clear and complete way that can be easily understood by learning experts and thus contribute to the enhancement of explicability.

5. Discussion and Perspectives

The presented results show that it is possible to define learner personas based on learning indicators computed from learning traces. The applied methodology enables to detect several outliers and then to regroup inliers into homogeneous subsets of students. All the students, whether they are inliers or outliers, are then described in the form of personas.

On the one hand, personas make it possible to represent a wide variety of behaviors adopted by the student population studied. It is to these different subsets of students that educational systems must be able to respond indiscriminately, even if some groups are representative of a larger or smaller population

of students. Personas representing a very small number of students, or a single student, deserve as much attention as others and should not be dismissed. That is why we talk about representativeness: all students, regardless of their behavior, must receive the help that is adapted to them, always with the same quality and without some being over-, under-, or non-represented. On the other hand, embodying the results of LA algorithms in the form of personas seems to us to be an important step towards improving the explicability of systems, and at the same time, we have good hopes for increasing user confidence, reaching a wider audience, and having a positive impact on various stakeholders. Overall, this study provides a new approach to evaluate SLEs in a fair way, based on explainable LA in order to increase user confidence while developing more ethical systems.

As a follow-up to this work, we plan to study some specific categories of learners, as repeated students, and to examine the presence of specific student profiles defined in the literature, such as those detailed in the ICAP model (23).

6. Acknowledgments

This work is done in the framework of the LOLA project, with the support of the French Ministry of Education. LOLA ("Laboratoire Ouvert de Learning Analytics") allows the evaluation of LA tools.

7. References

1. Siemens G, Long P. Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE Review. 2011;46(5):30.

2. Slade S, Prinsloo P. Learning Analytics: Ethical Issues and Dilemmas. American Behavioral Scientist. 2013;57(10):1510–29.

3. Drachsler H, Greller W. Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16. Edinburgh, United Kingdom: ACM Press; 2016.

4. Holmes W, Porayska-Pomsta K, Holstein K, Sutherland E, Baker T, Shum SB, et al. Ethics of AI in Education: Towards a Community-Wide Framework. Int J Artif Intell Educ; 2021.

5. Erdt M, Fernández A, Rensing C. Evaluating Recommender Systems for Technology Enhanced Learning: A Quantitative Survey. IEEE Transactions on Learning Technologies. 2015;8(4):326–44.

6. Iksal S. Ingénierie de l'observation basée sur la prescription en EIAH. 2012.

7. Ben Soussia A, Roussanaly A, Boyer A. An in-depth methodology to predict at-risk learners. 16th European Conference on Technology Enhanced Learning [Manuscript submitted for publication]. 2021.

8. Mupinga DM, Nora RT, Yaw DC. The Learning Styles, Expectations, and Needs of Online Students. College Teaching. 2006;54(1):185–9.

9. Xu D, Jaggars SS. Performance Gaps between Online and Face-to-Face Courses: Differences across Types of Students and Academic Subject Areas. The Journal of Higher Education. 2014;85(5):633–59.

10. Mojarad S, Essa A, Mojarad S, Baker R. Data-driven learner profiling based on clustering student behaviors: learning consistency, pace and effort. 2018.

11. Haiyang L, Wang Z, Benachour P, Tubman P. A Time Series Classification Method for Behaviour-Based Dropout Prediction. In: 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT). 2018. p. 191–5.

12. Tempelaar D, Rienties B, Mittelmeier J, Nguyen Q. Student profiling in a dispositional learning analytics application using formative assessment. Computers in Human Behavior. 2018;78:408–20.

13. Lotsari E, Verykios VS, Panagiotakopoulos C, Kalles D. A Learning Analytics Methodology for Student Profiling. In: Likas A, Blekas K, Kalles D, editors. Artificial Intelligence: Methods and Applications. Cham: Springer International Publishing; 2014.

14. Paiva ROA, Bittencourt II, da Silva AP, Isotani S, Jaques P. Improving pedagogical recommendations by classifying students according to their interactional behavior in a gamified learning environment. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. Salamanca Spain: ACM; 2015.

15. Kuzilek J, Hlosta M, Zdrahal Z. Open University Learning Analytics dataset. Sci Data. 2017;4(1):170171.

16. Hussain M, Zhu W, Zhang W, Abidi SMR. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computational Intelligence and Neuroscience; 2018.

17. Arnold KE, Pistilli MD. Course signals at Purdue: using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12. Vancouver, British Columbia, Canada: ACM Press; 2012.

18. Boroujeni MS, Sharma K, Kidziński Ł, Lucignano L, Dillenbourg P. How to Quantify Student's Regularity? In: Verbert K, Sharples M, Klobučar T, editors. Adaptive and Adaptable Learning. Cham: Springer International Publishing; 2016. p. 277–91.

19. Pluck G, Johnson HL. Stimulating curiosity to enhance learning. 2011;9.

20. Brooks C, Greer J. Explaining predictive models to learning specialists using personas. In: Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14. Indianapolis, Indiana: ACM Press; 2014.

21. Lallemand C, Gronier G. Méthodes de design UX: 30 méthodes fondamentales pour concevoir et évaluer les systèmes interactifs. Paris: Eyrolles; 2016.

22. Navarro ÁAM, Ger PM. Comparison of Clustering Algorithms for Learning Analytics with Educational Datasets. IJIMAI. 2018;5(2):9–16.

23. Chi MTH, Wylie R. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. Educational Psychologist. 2014;49(4):219–43.