École Polytechnique Fédérale de Lausanne

Exploring Learning Profiles: Time Series Clustering Guided vs. Free Training

CS-421 Machine Learning for Behavioural Data

Team BD4ED: Tobias Oberdörfer, Maxime Lelièvre, Violeta Vicente Cantero

June 8, 2023

***Abstract*—Over the last few years, the number of educational platforms has risen exponentially. Among them, Calcularis is a mathematics learning program for children from 1st grade to high school that offers a multi-sensory learning experience. This report presents a time series clustering approach to identify different types of learning profiles and compare them between the two training modes the platform offers. The results suggest that the choice of the features used is a key determinant of the clustering success.**

I. Introduction

Calcularis is a mathematics learning program for children from 1st grade to high school. It proposes a multi-sensory learning experience through a lot of different games which create a vast quantity of data that can be analyzed. A particularity of this platform is that the children can either learn via guided training, where they are put into games, or free training, where they can choose themselves what to play. While some children better learn when they are being guided, some others do not need guidance to be fully engaged with the task. In fact, it can even be counter-productive for their learning to be forced into specific tasks. Besides that, the dynamic nature of learning makes it hard, as well as critical, for a learning platform like Calcularis to identify which type of training better corresponds to a given user at a given time. Understanding the learning profiles of individuals is crucial for optimizing educational strategies and designing personalized learning experiences. Hence, by identifying different types of learners and comparing their profiles between guided and free training, we should be able to gain insights into the effectiveness of these two different approaches for different types of learners.

In this context, we investigate to what extent time series clustering can be used to identify different types of learners and categorize them into different learning profiles allowing a comparison between guided and free training. To this end, we transform the users' data into time series, train a clustering model on the data at hand and compare the resulting clusters of the guided training and free training.

After a description of the data and the exploratory analysis that has been performed, the proposed approach is explained, and the experimental evaluation, before discussing our final results and their implications.

II. Data analysis

A. Data description

The Calcularis dataset consists of three main data frames: *users*, *events*, and *subtasks*. The full dataset has 64 932 users, 2 185 200 events, and 3 502 884 subtasks. Their attributes are presented in Table I.

Table	*Attributes*	*Description*
Users	user_id	Unique id of user in database
	learning_time_ms	Time spent learning (ms)
	start	Timestamp of initial login
	end	Timestamp of last log-off
	logged_in_time_ms	Total time between login and log-off
	language	User's studying language
	country	User's country
Events	events_id	Unique id of event in database
	user_id	Unique id of user in database
	type	Type of event
	mode	Type of game played
	game_name	Game's name
	learning_time_ms	Total learning time over a single game (ms)
	number_range	Exercise's difficulty level
	start	Start timestamp of the event
	end	End timestamp of the event
	skill_id	Event's skill number
Subtasks	subtask_id	Unique subtask's id in database
	event_id	Unique event's id in database
	user_id	Unique user's id in database
	answer	User's answer
	correct	Boolean denoting answer's correctness
	correctAnswerObject	Correct answer
	hasProperResult	Boolean denoting the question is answerable
	range	Exercise's difficulty level
	subtask_finished _timestamp	Subtask's ending timestamp
	type	Subtask's type

TABLE I: Attributes of the datasets

While using the program, each user plays a game associated with a skill to learn, which represents an event. Those events have different subtasks, each of them with its difficulty level and correctness values, which are used to identify the user's ***Assessment***. The platform's nature and the lack of monitoring information about users' interaction with the platform heavily conditioned feature extraction, whose main goal was to analyze different dimensions related to self-regulated learning. For instance, information such as the number of clicks the user does while solving a question or how many times the user corrects their answer before submitting is interesting information that would have been used to extract ***Consistency*** or ***Effort*** features. On the other hand, ***Proactivity*** features such as content anticipation cannot be extracted from the given data, as its unknown when the user first encountered that specific skill, as that could have been in school or anywhere else outside of the Calcularis platform.

Given our interest lies in the comparison of the behavior between the students that are doing guided training and those doing free training, ideally we would have the same amount of data points for both. This is not the case in the given Calcularis data set though. As can be seen in Figure 1, almost all events in our dataset are purely between the NORMAL and END_OF_NR modes, which both count towards guided training.

B. Exploratory analysis

The exploratory analysis phase of our study involved at first an examination of the given dataset, which was

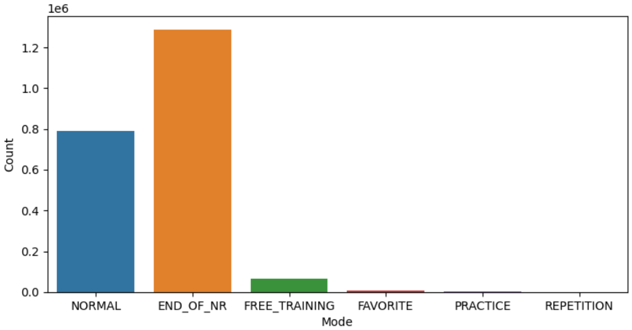


Fig. 1: Distribution of game modes in the *event* data frame.

completely unknown to us. We began by investigating the distributions of various features within the dataset, including the number of different game modes available and the different skills within the training platform. We did so both through the dataset itself, but also through the provided access to the platform using the test account. Understanding the meaning and purpose of each mode was crucial for ideating diverse research questions that could improve the learning experiences offered to students.

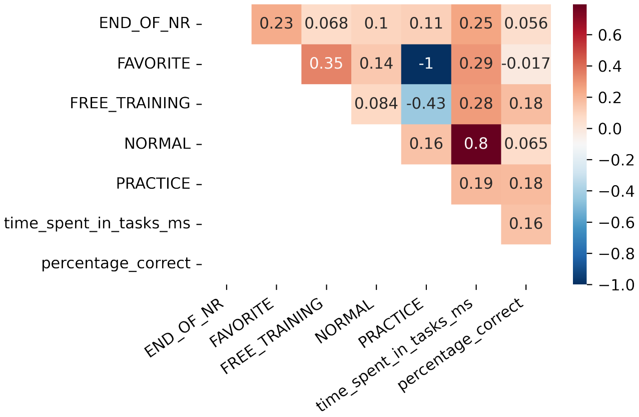


Fig. 2: Pearson's correlation between learning time in milliseconds of the different game modes and the percentage correct over a student.

Pearson's correlation between learning time in the different game modes for a small dataset of 1000 users was explored. Additionally, potential correlations between the game modes and the average correct percentage achieved by the learners were studied, aiming to identify relationships between these two metrics. As seen in Figure 2 there seem to be high correlations between the different gaming modes, but the most interesting part, which also had an influence on our research question is that there seems to be a correlation between both practice time and free training time spent with the total percentage correct a student achieved across their time on Calcularis.

A study about the correlation between the number of games played per user and the answer's correctness was also performed. An example of such studies can be seen in Figure 3, showing how higher numbers of games played, result in higher percentage correct; meaning that users who play more games achieve a higher number of correct answers. Realizing that characterizing students by their performance and studying tendencies was a shared interest led to the final research question.

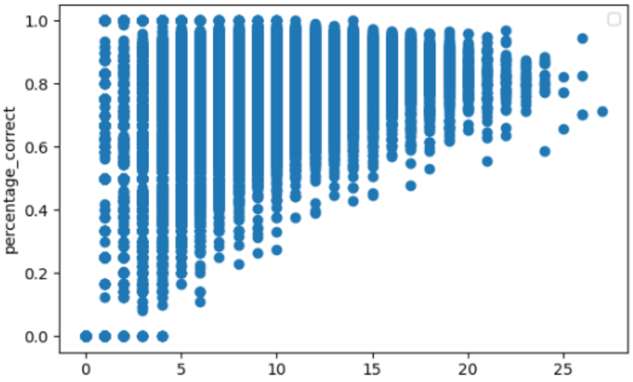


Fig. 3: Percentage of correct answers depending on the number of games played.

Lastly, while analyzing these aspects of the dataset, we also considered the possibility of incorporating the skill map. However, due to its substantial size and the intricate nature of the relationships of the flow along the graph, we recognized that its inclusion would require more extensive research, beyond our scope.

Looking back at our exploratory analysis, we now acknowledge that at the time, we did not fully consider the concept of fairness or rather the concepts we learned with fairness, as it pertains to learner profiling. In hindsight, integrating fairness considerations into our analysis could have been a valuable approach to reach any understanding of the potential disparities and biases within the dataset, other than that most students come from Switzerland. Examining the dataset through a fairness lens would have enabled us to assess whether the training modes disproportionately favor certain groups or if there are any unintended consequences for specific learner demographics. Despite this missed opportunity, our initial exploratory analysis provided important insights and shaped the trajectory of our project.

III. Proposed Approach

A. Data preprocessing

The first step is to split all *events* and *subtasks* into guided and free training, such that later a comparison between them is possible. To this end, the *mode* in the *events* data frame is selected. Specifically, guided and free training takes all the events and subtasks that have the mode NORMAL, END_OF_NR, FAVORITE, REPETITION, and FREE_TRAINING, respectively. This first step results in having 59817 different users for the guided training and 8137 for the free training, which counts for 88% and 12% of the total users respectively (see Figure 4).

In addition to this heavy imbalance in the number of users in the two training modes, the imbalance is even more prominent for the number of games played in each game, with 97%

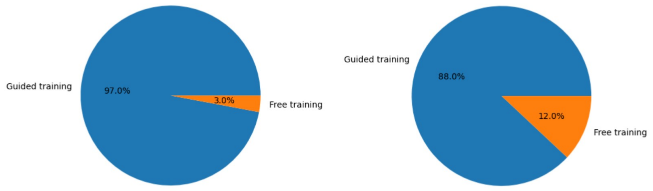


Fig. 4: Guided training (blue) vs free training (orange): number of games on the left; number of users on the right.

and 3% for guided and free training respectively. Though the difference in the number of users and games in each training mode is important, the chosen approach consists in clustering a subsample of guided training, with 8 000 users actually, for computational reason. Another alternative could have been to cluster on a balanced combination of guided and free training, for instance randomly selecting 8 000 users in guided and free training separately and feeding the clustering model with those 16 000 users. However, one of the goals of the research question is the comparison between the learning profiles between the two training modes, we thought that clustering on a balanced dataset could lead to a lot of overlaps and make the comparison harder between guided and free training. The clustering model is thus trained only with guided training, a subsample for computational reasons, which will later be used to compare with free training. Indeed, if there are some common learning profiles between the two training modes, the clusters should be identifiable with the free training data.

The second step of preprocessing is the *add_week_dimension* function, which introduces the time-related information by using the *start* and *subtask_finished_timestamp* and setting this new dimension to 0 for the first week a user started using Calcularis.

The functions *effort_features* and *assessment_features* build features per user and per week. Those features aim to combine multiple behavioral dimensions to obtain interpretable student profiles and most of them are inspired by the paper ”Identifying And Comparing Multi-Dimensional Student Profiles Across Flipped Classrooms” [1].

The dimension ***Effort*** aims to track the level of student involvement in the course while ***Assessment*** aims to track how much the users master the different skills.

The ***Assessment*** feature has been created from the attribute *number_range*, which has 4 different values; R10, R20, R100, R1000. In order to get a better understanding of the feature, a *default* and *linear* scaling of the *number_range* attribute was used, which helps with interpretability. The *default* scaling states that a game with a *number_range* of R10 is twice as easy as a R20 game, and 100 times easier than a R1000 game. The *linear* scaling states that a game with a *number_range* of R10 is twice as easy as a R20 game, and 4 times easier than a R1000 game. In the end, the *linear* scaling has been chosen for realistic reasons, a game 100 times more difficult did not really make sense.

Dim.	*Feature*	*Description*
Eff.	time_online	Total time spent online
	events_done	Total number of games played
	subtasks_done	Total number of subtasks done
	Nb_of_different_games	Number different games played
	correct	Percentage of correct answers
Asses.	range_linear	Number range of each event
	user_num_attempts	User attempts per event
	num_users_tried	Number users tried each event
	attempts_diff	Relative number of event attempts
	min_range_lin	Lowest event’s difficulty per user
	avg_range_lin	Average event’s difficulty per user
	max_range_lin	Highest event’s difficulty per user
Reg.	week_day_periodicity	Studying on certain day(s)
	week_hour_periodicity	Studying during certain day hours
	day_hour_periodicity	Studying on certain week hours

TABLE II: Features are grouped into different dimensions.

The *create_smaller_dataset* function does the subsampling of students to be able to run the models. It randomly selects a desired number of users from the *events* data frame and takes the corresponding subtasks.

Among all the features mentioned in Table II, the chosen features for the clustering are *time_online*, *events_done*, *subtasks_done*, *Nb_of_different_games*, *correct*, *attempts_diff* and *avg_range_lin*. The regularity features are not used because they are calculated per student over the whole time and not per week. Calculating them per week is not possible for most students because there is not enough data per student per day to calculate them with relevance. The different features have different scales, standardization is applied to all features to ensure every feature contributes the same amount to the clustering. This combination of features is then reshaped as time series to become the inputs of the clustering models. A lot of users describe inconsistent activity on the platform, thus the time series obtained needs to be filled everywhere there are missing values (weeks without any activity).

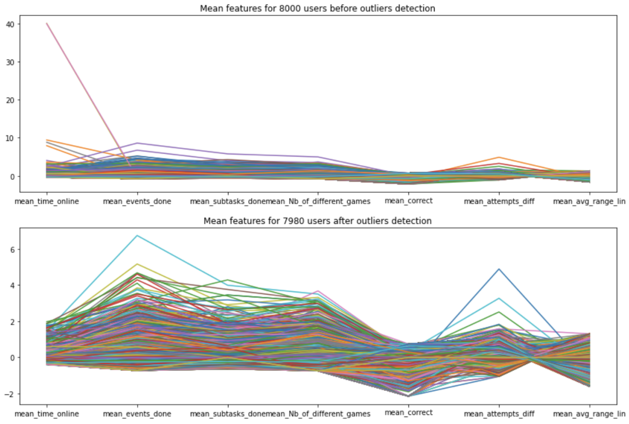


Fig. 5: Mean lines per user and per feature before (up) and after (down) outliers detection.

Prior analysis ended up with clusters with only one user in it, suggesting big outliers. Thus, after combining the used features together, the users detected as outliers are removed. To this end, the average values over the first 16 weeks per user and per feature are extracted and the users with an average value further than 3σ from the mean in at least one feature are removed. The choice of 3σ comes as a trade-off between removing extreme values before creating the clusters and keeping as many users as possible. First, it is worth noting that even after the standardization, there remain extreme values that could lead to clusters having still only single users. Second, even though some features do not look to have outliers first as shown in the first plot of Figure 5, some experiments showed that it really depends on the subsample taken. Outliers detection only on the *time_online* feature was found to achieve the best silhouette score, probably because this feature has the biggest outliers compared to all others. For the 8k subsample of guided training used for clustering, 20 outliers have been removed.

B. Models and Methods

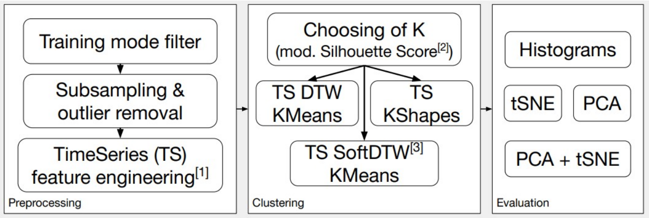


Fig. 6: Model architecture

The initial model to cluster the students (not shown in the handed-in Jupyter notebook, but can be found here) was spectral clustering using a dimensionality matrix.

This implementation was heavily based on the paper by P. Mejia-Domenzai et al. [1]. and hence has many of the same features. Among the six different dimensions in which the features were originally grouped, our analysis focuses on the ***Effort*** and ***Assessment*** features. In the aforementioned paper, the first clustering step is done separately per dimension by first creating a similarity matrix using Dynamic Time Warping, using a Gaussian kernel to transform the distance matrix into a similarity matrix, and finally adding the similarity matrices of the features of each dimension to get the dimension similarity. After that, spectral clustering is used to cluster the similarity matrix of each dimension separately. As a result, this initial implementation of spectral clustering is infeasible for the full dataset. The idea of doing spectral clustering on a sensible subset of the dataset, like a specific country or set of games, to reduce the number of students was quickly discarded. The reason is, that any such subset would still contain students in the thousands, which cannot reasonably be run in time with our hardware. This leads us to change our design from spectral clustering using similarity matrices to using direct time-series K-Means on our features.

Despite not being able to use the implementation of spectral clustering by P. Mejia-Domenzai et al. [1], it was still possible to reuse many of the same time-series features. We explored the utilization of K-means clustering, employing both DTW (Dynamic Time Warping) and softDTW [2] distances, as well as K-Shape. To determine the appropriate number of clusters (K) a modified silhouette score (SS) has been devised, which takes into account the computational complexity of the algorithm. This function computes several times a silhouette score on a random stratified subsample and takes the average silhouette score over all the random subsamples. In the end, the clustering has been done with softDTW because it gave the best silhouette score compared to DTW and K-Shape.

The models created, the last challenge was to be able to explain the results. As for most machine learning approaches, it is a particularly difficult and critical part because of the multi-dimensionality of the used features and all the transformation the raw data underwent before feeding the models. To this end, the clusters' barycenters of each feature are plotted to manually infer what the clusters mean. These graphs allow us to see the different users' profile distributions over time for each cluster, compare them, and have some insights into what defines a particular cluster. The plots of the barycenters mostly allow us to explain the clusters' meaning but are limited because of the separability of the clusters, even though the silhouette scores are decent.

To evaluate the separability in more detail PCA and PCA+t-SNE plots are used for their effectiveness in visualizing high-dimensional data. t-SNE plots are known to be hard to interpret although they can give meaningful information about the separability of the clusters. Tuning the perplexity hyperparameter can help, but it seems our clusters are just not well separable.

IV. Experimental evaluation

Results achieved initially with the spectral clustering, for a few hundred students, are not significant, because they were built with such a small data subset. Although it even took the order of hours to get the results for these, there was no clear number of clusters that could have been taken from between two and nine clusters. The silhouette score plot only showed decreasing scores for more clusters. The distortion plot shows that three clusters are optimal, but finding an interpretation of those clusters was not possible.

Because time-series K-means does not have the same explosive growth in the number of students as producing a dimensionality matrix, TimeSeriesKMeans managed to cluster our features on a subset of 25k guided-training students within a few hours. Even still, all of the results are shown on an 8k sub-sample to have the same sized sets between guided and free training. The SoftDTW metric was chosen over DTW based on the average Silhouette score over all different numbers of clusters (seen here). The final $K = 6$ was chosen on a subset of 9.5k guided training students.

After clustering with the before-described architecture, the cluster membership numbers seen in Figure 7 are achieved. This clearly shows two things; First, cluster number zero has overwhelmingly many students for both guided and free training, with 71% and 87.5% respectively. Secondly, clusters one and two do not show up in the guided training pie chart, whereas clusters one and five do not in the free

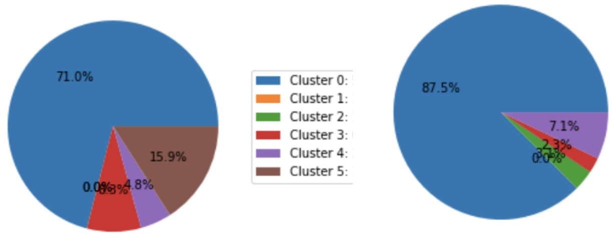
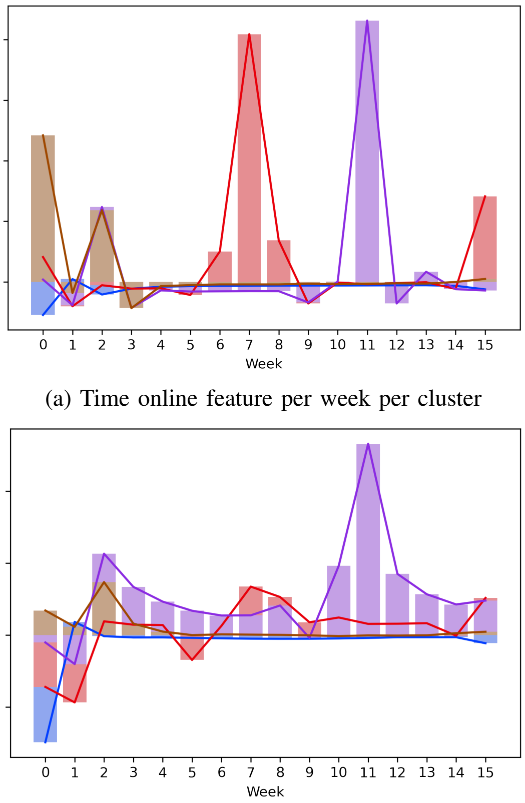


Fig. 7: Left: 7980 guided users with a silhouette score of 0.269 (Std. 0.026); Right: 8000 free training users with a silhouette score of 0.474 (Std. 0.042)

training chart. In actuality, both of the pie charts do have those clusters as well, but because they only have up to two students in those clusters they are not visible. Already from this cluster distribution, it is clear that our clustering is not great, given that there are even clusters with just single-digit amounts of students. One possibility for these cluster distributions might be that the outlier detection did not remove all outliers or rather removed the extreme outliers but the rest of the dataset still contains more. As such cluster one and two seems to contain outliers for guided training, whereas cluster one and five are the outlier clusters for free training.



(a) Time online feature per week per cluster

(b) Correct answers feature per week per cluster

Fig. 8: Barycenters plotted for all clusters with more than two students and two different features.

The barycenters of the four non-outlier clusters for guided training and the two features *time_online* and *correct* can be seen in Figure 8. Those two features are shown for two reasons; The barycenter for *time_online* looks very similar to many of the other effort features like *subtasks_done* and *Nb_of_different_games*. This makes sense considering doing any task, event, or game requires the student to invest more effort. The second reason is that the *correct* feature is what makes the clusters interesting in the first place, given the ideal goal in education of helping all students learn in the most efficient way.

In general, the clusters seem to assign different types of learners to the different clusters. For example, clusters one and five have very different distributions of the percentage correct over the 16 weeks compared to clusters three and four. Both clusters three and four in Figure 8a seem to correspond to students that have bursts of activity at some point after the first few weeks. When additionally looking at the feature of *time_online*, the cluster of students that invest comparatively more time in training, i.e. clusters three and four, generally also have a higher percentage of correct answers towards the end. Although there are outlier clusters and cluster zero for free training is much bigger, clusters three and four can be seen to differ strongly from the big set of cluster zero in Figure 8b.

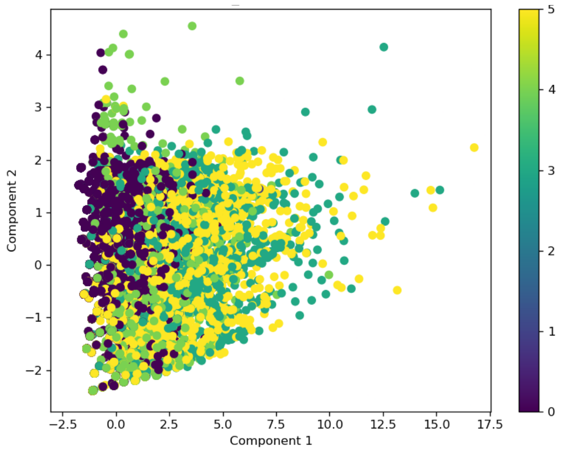


Fig. 9: 2D PCA plot for the guided training clusters.

Plotting the result of a principal component analysis (PCA) is one way to evaluate the goodness of clustering. PCA reduces the high-dimensional clusters to their two principal components, which are plotted in Figure 9 for the guided training clusters. For the six different clusters, we see the color assignment on the right side. Looking at the plot we can clearly see that the clusters are not very well separated, i.e. our clustering approach did not yield well-separated clusters. Still, it is understandable from the plot that not all clusters are completely overlapping, underlining the achieved silhouette score. Specifically, clusters zero and one seem to be clustered rather closely together, whereas both clusters four and five seem to be more spread out but somewhat separated from cluster zero.

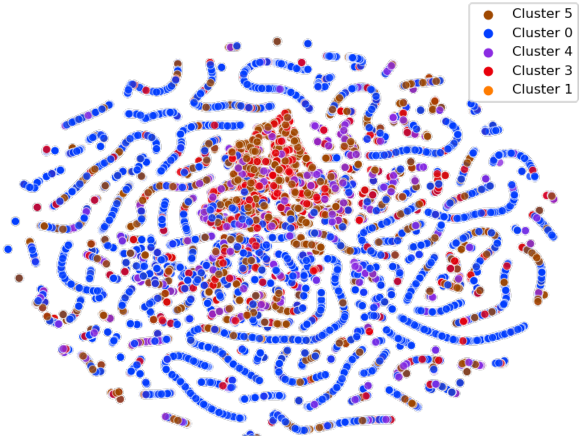


Fig. 10: Cluster separation visualization using PCA + tSNE for the guided training clusters.

Another way to evaluate cluster separation is a tSNE plot. Figure 10 shows the tSNE plot of the same guided training clusters, which were first reduced in dimension using PCA. Again, similar to the pure PCA plot, it is clear that our clustering approach did not yield very well-separated clusters, but the tSNE shows much more clearly that big parts of the clusters are separable. Specifically, cluster five and big parts of cluster three are overlapping, and cluster zero shares some overlap with them, but not much.

V. Discussion and implications

The presence of many overlapping clusters of learners suggests our features may not be sufficient to capture the full spectrum of learner characteristics. One potential reason could be the complexity of learning behaviors. Another might be there are simply too many different learner types because nearly every student could be unique. Moreover, it is possible that the distinction between guided and free training is not as clear-cut as initially anticipated.

Still, the identification of distinct learner profiles might inform future instructional design and personalized learning approaches of Calcularis and align the employed teaching strategies better with the diverse needs of their user base. Additionally, suggested skills or game modes could be customized, ensuring that learners receive the incentives and interventions that are most effective for their particular learner profile.

When interpreting our results several encountered limitations and challenges need to be considered. Firstly, our approach of comparing learner types between guided and free training modes has inherent limitations because the data points from guided training are not independent under the influence of the guiding algorithm. This raises the possibility that the observed clusters in our results might actually represent internal clusters of the guiding algorithm. Secondly, the comparison between students' free and guided learning behavior may not fully reflect their real-world behaviors, as both cases are constrained within the possible action space of Calcularis. Additionally, detection and handling of outliers presented a challenge, because it is difficult to know if all outliers were removed before the clustering process.

Moreover, computational constraints were hit for spectral clustering, which is known to work well for such a use-case, and the normal silhouette score due to the size of our dataset, necessitating alternative approaches. Another limitation, which we, unfortunately, did not account for by neglecting stratified sampling, lies in the imbalance of the given data, with approximately 90% of the data originating from guided training. Lastly, free training profiles may not be directly comparable to guided training profiles, as they might represent completely different learners.

Because of the above-mentioned, our research question can only be partially answered. Although different learning profiles were identified, as illustrated in Figure 8, it is crucial to acknowledge that our results did not yield a satisfactory answer. The overlapping clusters show our utilized features do not categorize learners well in guided and free training, let alone allow an actionable comparison between them.

VI. Conclusions and future work

In summary, our project revealed a crucial finding with some implications for Calcularis. While only the first part of our research question was answered, as we were only able to partially identify different learners, additional features or further modifications to our architecture are necessary. The key takeaway is the need for exploring additional data sources and features to enhance the accuracy of learner profiling. By acknowledging this finding and considering potential modifications, Calcularis can strive towards a more comprehensive understanding of learner characteristics, leading to better-personalized learning experiences for their users.

Despite the challenges encountered and the overlapping clusters observed, this pipeline for identifying learner types in the context of guided versus free training modes serves as a basis for future investigations. By documenting the methodology, preprocessing steps, and clustering techniques employed, as well as those that are not feasible, a road map to build upon is provided, offering a starting point for further work. The authors hope this pipeline helps in the further development of personalized learning experiences with the Calcularis application.

Any future work could start with improvements and modifications to our model architecture, building upon the identified limitations and challenges. To achieve this, exploring alternative clustering algorithms that can better handle the complexities of the data could be beneficial. But this could also include going back to the spectral clustering approach with more computing power on hand. Additionally, refining the feature engineering process by incorporating additional relevant variables or exploring more advanced techniques of clustering might be helpful.

To further enhance the accuracy of learner profiling for Calcularis, there is great potential in incorporating additional data currently not captured. While our focus was on time series data derived from guided and free training games, integrating supplementary information could provide a deeper understanding of learner characteristics. For instance, incorporating demographic data, such as age, gender, or educational and monetary background, may shed light on how these factors influence student learning types. When adding features based on such demographic data it is crucial to make sure they do not also introduce biases. This could be checked with metrics like equalized odds, demographic parity, or similar other metrics. Furthermore, self-reported motivation, engagement levels, or cognitive abilities could provide valuable information on the underlying factors that contribute to the identified clusters. This broader scope of data collection plays into the idea of needing a more diverse range of features for a more nuanced characterization of learner profiles.

Another avenue for future work could be the incorporation of interviews or surveys to gather subjective insights from the assigned students for the identified learner types to figure out if the assignments or general clusters make sense. These insights of both students and teachers could be especially helpful to guide any architecture or feature improvements in a more goal-oriented fashion.

Further complimentary research could look at how students in the identified learner profiles evolve between them. Maybe some specific learners have a much lower probability of advancing towards a ”better” learner profile and could benefit from interventions towards them.

Acknowledgements

The authors want to thank Paola Mejia-Domenzain for access to the flipped classroom source code and for all the great insights into clustering, as well as Vinitra Swamy for the quick and effective solutions to our multiple concerns about the data and the project. Further, thanks go out to all of the projects which were used during the course of this project, especially the Tslearn [3] time series library, without which this work would have not been possible in the given time.

References

[1] P. Mejia-Domenzain, M. Marras, C. Giang, and T. Käser, “Identifying and comparing multi-dimensional student profiles across flipped classrooms,” in *Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 90–102. [Online]. Available: https://doi.org/10.1007/978-3-031-11644-5_8

[2] M. Cuturi and M. Blondel, “Soft-DTW: A Differentiable Loss Function for Time-Series,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML’17. JMLR.org, 2017, p. 894–903.

[3] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, “Tslearn, a machine learning toolkit for time series data,” *Journal of Machine Learning Research*, vol. 21, no. 118, pp. 1–6, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-091.html