CS-421 Class Project 2023
Team BD4ED

Tobias Oberdörfer Maxime Lelièvre Violeta Vicente Cantero

I. Introduction

The aim of the project is to use time series clustering per student to look at the difference between guided training and free training, with the final goal of learning what type of learner each student is and potentially be able to categorize them between different students' profiles.

II. Data analysis

A. Data collection

The dataset at hand comes from Calcularis, a mathematics learning program, and consists of three main tables: *users*, *events* and *subtasks*. The full dataset has 64 932 users, 2 185 200 events and 3 502 884 subtasks. While using the program, each user plays a game associated to a skill to learn, which represents an event. Some games can have different subtasks.

B. Data preprocessing

Events and *subtasks* tables are split between the guided and free training. For this the *mode* in the *events* dataframe is selected. Specifically, guided and free training take 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. Though the difference of the number of users in each training mode is important, our clustering approach uses a subsample for guided training to decrease the difference

The add_week_dimension function 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.

The functions *effort_features* and *assessment_features* build features per user and per week. Those features will later be reshaped as time series to become the inputs of the clustering models. Those features aim to combine multiple behavioral dimensions to obtain interpretable student profiles and most of them are inspired from 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 ***Regularity*** feature is currently not in time-series format and hence not used anymore.

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 *range* attribute was used, which helps with interpretability.

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 I
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* dataframe and takes the corresponding ones in the *subtasks* dataframe.

III. Models and Methods

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 and hence has many of the same features. Among the six different dimensions in which the features were originally grouped, our focused is on the ***Effort***, ***Regularity*** and ***Assessment*** feature. In the aforementioned paper, the first clustering step is done separately per dimension by first creating a similiarity 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. However, due to the much larger number of students in the Calcularis dataset (roughly 65k students), compared to the flipped classroom dataset, it would take on the order of 60 days to create the similiarity matrix. 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 amount of students was quickly discarded. The reason being, that any such subset would still contain students

in the thousands, which cannot reasonably be run in time with our hardware. This lead us to change our design from spectral clustering using similarity matrices to using directly time-series K-Means on our features.

Despite not being able to use the implementation of spectral clustering by P. Mejia-Domenzai et al, it was still possible to reuse many of the same time-series features. The current implementation uses tslearn's[2] TimeSeriesKMeans. Out of the possible three clustering metrics (Euclidean, DTW and SoftDTW [3]), which TimeSeriesKMeans supports, the Euclidian distance was directly excluded because the students in the Calcularis dataset do not work at the same time on the same problems (neither in guided training nor free training).

Plotting the clusters' baricenters is a nice way to manually infer what the cluster mean and detect outliers when there is presence of 0-n clusters. These graphs allow us to see the different users' profiles distributions over time.

IV. Results

The initial results of the spectral clustering, for a few hundred students, are not significant, because they were build with such a small data subset. Although it even took in 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 score 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 for all free-training students and a subset of 25k guided-training students. 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 because the Silhouette score has similar computational complexity to a similarity matrix.

Cluster index	*Guided training*	*Free training*
0	24995	6485
1	1	360
2	1	157
3	1	2
4	1	1129
5	1	4

Table II
Number of students per cluster for guided and free training.

As can be seen in the table above there clearly are clusters that contain outliers (one through five for guided training, three and five for free training). The distribution across the other clusters seems more reflective of actual behavioural clusters students could be assigned to, still these clusters will change when removing outliers first.

The baricenters of the four non-outlier clusters for free training and the feature *correct* can be seen in Figure 1. In general the clusters seem to assign different types of

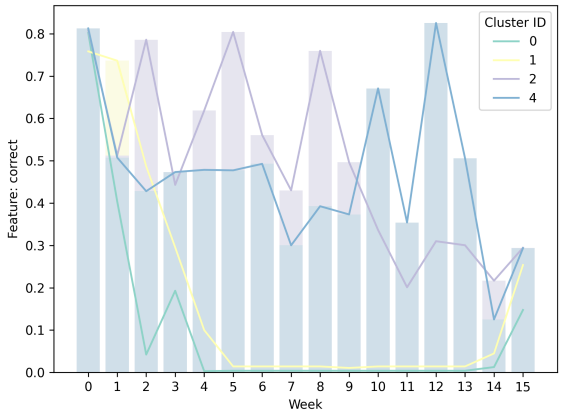

Figure 1. Non-outlier baricenters of free training student clusters.

learners to the different clusters, as such for example one can see that clusters two and four above have a very different distribution of the percentage correct over the 15 weeks looked at. When additionally looking at the feature of *time_online*, the cluster of students that invest more time in training generally also have a higher percentage of correct answers over their learning time. Although there is outlier clusters and cluster zero for free training is much bigger, cluster two and four can be seen to differ strongly from the big set of cluster zero in Figure 1.

Our approach of comparing types of learners between the two different training modes in Calcularis has two main limitations. First, the data points of the guided training are by no means independent, because all of the guided data points are influenced by the guiding algorithm. One consequences of this might be, that the recognizable clusters from our results are in fact internal clusters of the guiding algorithm, which were reconstructed from the students that went through guided training. The second limitation is that a comparison between student's free learning behaviour and their guided learning behaviour within Calcularis might not reflect their real behaviours, because in both cases they are still restricted to actions within the possible action space of Calcularis. A third limitation also appears when outliers are detected in the data, as this inherently makes it very difficult to be fully sure that all outliers have been found and smartly treated for the clustering.

The limitations of our approach being explained, our research question can only be partially answered. As shown in Figure 1, different learning profiles have been identified for the free training data. But because of the outliers that exist in the guided training data, relevant profiles could not have been identified for this data. Hence, only the second half of our research question is answered as of yet, namely that learning types can be identified.

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 to Vinitra Swamy for the quick and effective solutions to our multiple concerns about the data and the project.

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] 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

[3] 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.