EPFL

Machine Learning for Education Laboratory
ML4ED

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

On the Challenges of Identifying Learning Types in Calcularis

Tobias Oberdörfer, Maxime Lelièvre, Violeta Vicente Cantero
Machine Learning for Behavioral Data (MLBD) Course, EPFL 2023

1. INTRODUCTION + RESEARCH QUESTION

Purpose: Identifying types of learners in guided versus free training.

Method: Multivariate K-Means time series clustering.

Research question: Can time series clustering be used to identify types of learners and categorize them into different student profiles allowing for a comparison between guided and free training?

Key takeaway: The identified clusters were largely overlapping, suggesting the used features may not be sufficient. While the current approach may not have been successful, a working pipeline for future research to build upon was created.

· why K means? which one didn't work
bc of computational constraints

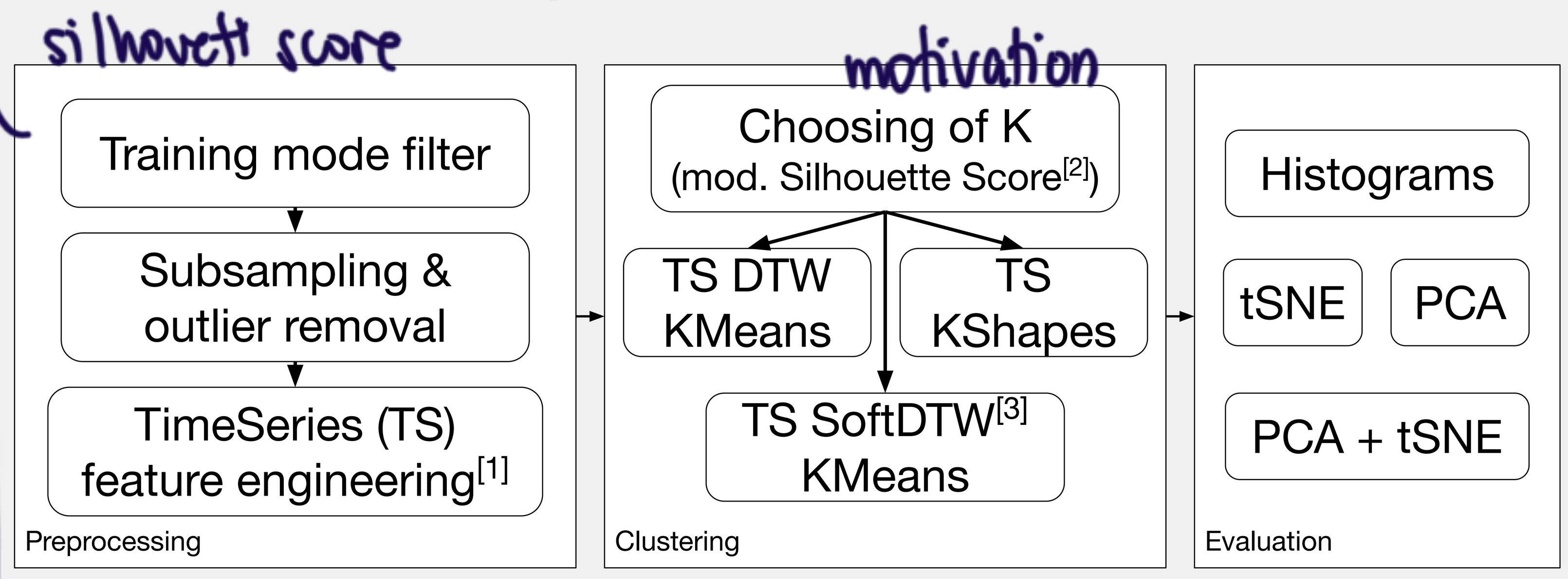
2. METHODOLOGY

- k is chosen to be 6 based on a modified silhouett score

+ modified silh. score

Model Architecture

motivation



Dataset

Engineered features
time_online
events_done
subtasks_done
Nb_of_different_games
correct
attempts_diff

Strategies to cope with the exponential computational cost of the models & metrics:

- Subsampling of the guided training dataset to train the model on 10 000 users only.
- Subsampling silhouette score averaged over multiple different runs.
- Removing all outliers based on three standard deviations from the mean.

• Feature engineering

Figure 1. Guided training vs free training: games and user number comparison.

• List features and invite listeners to ask about more concreteness

3. RESULTS

distribution of the clustering

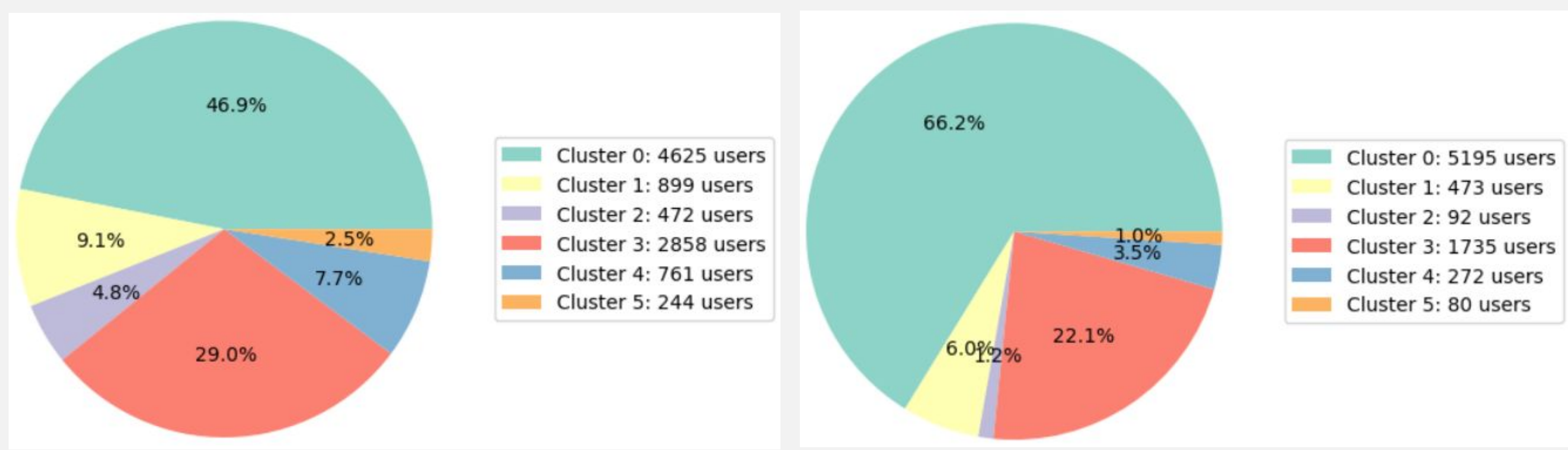


Figure 2. Left: Guided training with 9859 users sample and a silhouette score of 0.184
Right: Free training with 7847 users sample and a silhouette score of 0.386

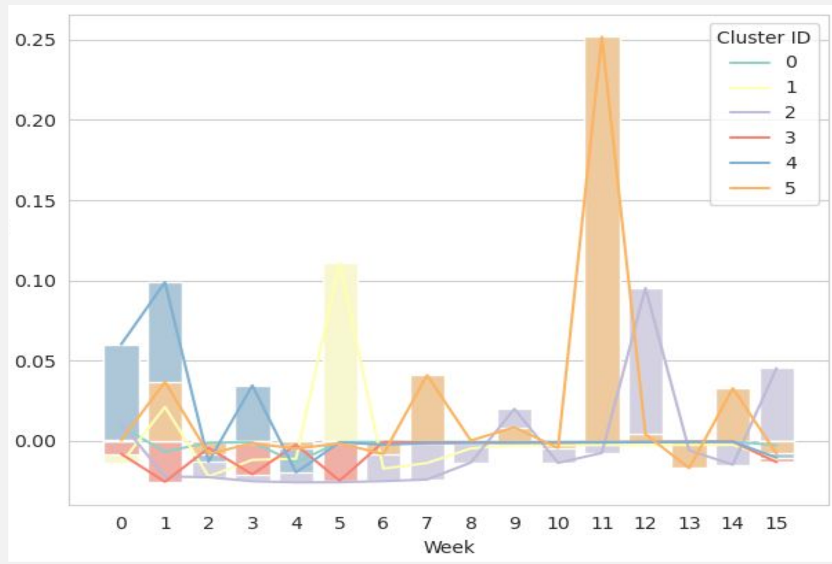


Figure 3. Plot of the online time feature cluster centroids for 16 weeks.

SoftDTW cluster centroids show different types of learners:

- Cluster zero clusters users that do not work much across all 16 weeks.
- In clusters three and four users seem to stop working after only a few weeks.
- Clusters two and five seem to work more between weeks 10 and 15 instead of more initially.

2&5: low at beg. & inconsistent at end

• cluster 0 = avg behavior
→ ⊕ diverse behaviors in GT
• 3 & 5: special early behavior (⊖, ⊖)

Figure 4a. Guided training

Figure 4b. Free training

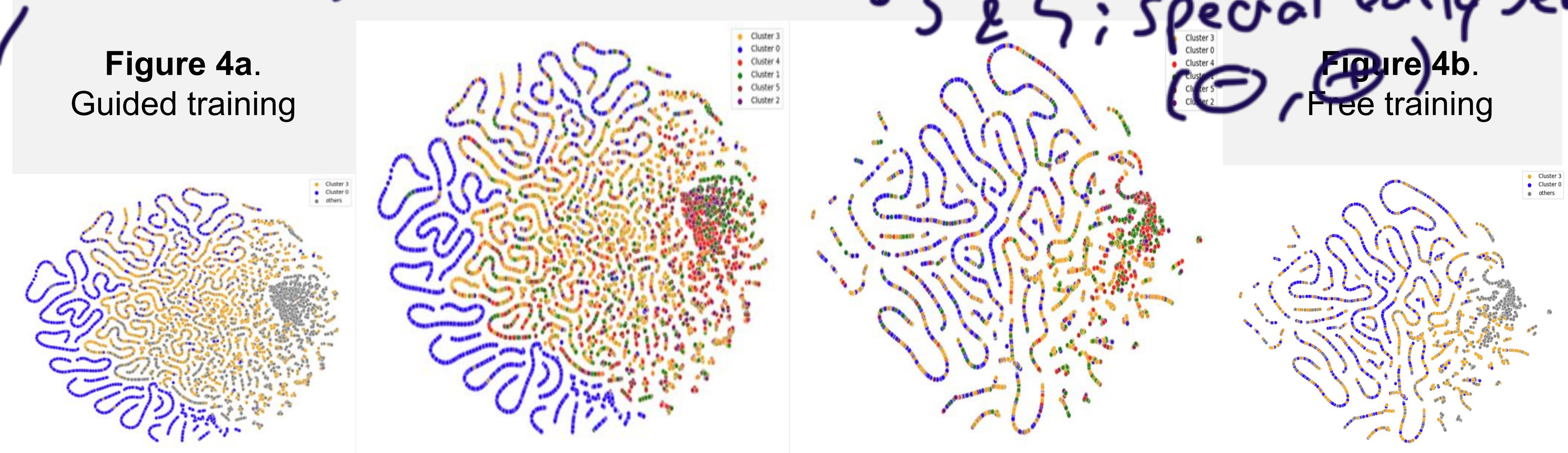


Figure 4. t-SNE plots of our clusters for different types of learners. *Middle:* visualization of all six clusters, *Outsides:* visualization combining overlapping clusters.

• describe features

clusters 0 & 3 = the less overlapping
↳ specially seen for outside tsne

4. CONCLUSION

- The **used features** for time series clustering **may not be sufficient** for identifying different types of learners.
- However, the study **provides a usable pipeline** to rerun the clustering approach with additional features in the future.
- Although **identified clusters overlap**, some **broad patterns** in student learning were **still recognizable**, suggesting that further exploration and feature refinement may lead to better differentiation.

REFERENCES

[1] P. Mejia-Domenzain et al., "Identifying and comparing multi-dimensional student profiles across flipped classrooms," in Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, Proceedings, Part I. Berlin, Heidelberg: Springer-Verlag, 2022, p. 90–102. [Online].

[2] R. Tavenard et al., "Tslearn, a machine learning toolkit for time series data," Journal of Machine Learning Research, vol. 21, no. 118, pp. 1–6, 2020. [Online].

[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.

École polytechnique fédérale de Lausanne