Accelerating Cluster Assignment for SeqClu

Rami Al-Obaidi / MSc. Azqa Nadeem & Dr. Ir. Sicco Verwer / Faculty of Electrical Engineering, Mathematics and Computer Science / Delft University of Technology

BACKGROUND INFORMATION

- SeqClu is a real-time sequence clustering using k-medoids algorithm.
- It uses 5 prototypes to represent a cluster.
- During cluster assignment phase, an incoming sequence is assigned to the cluster with lowest average distance between its prototypes and the incoming sequence.
- SeqClu uses Dynamic Time Warping (DTW) as distance metric instead of Euclidean distance.

RESEARCH QUESTION

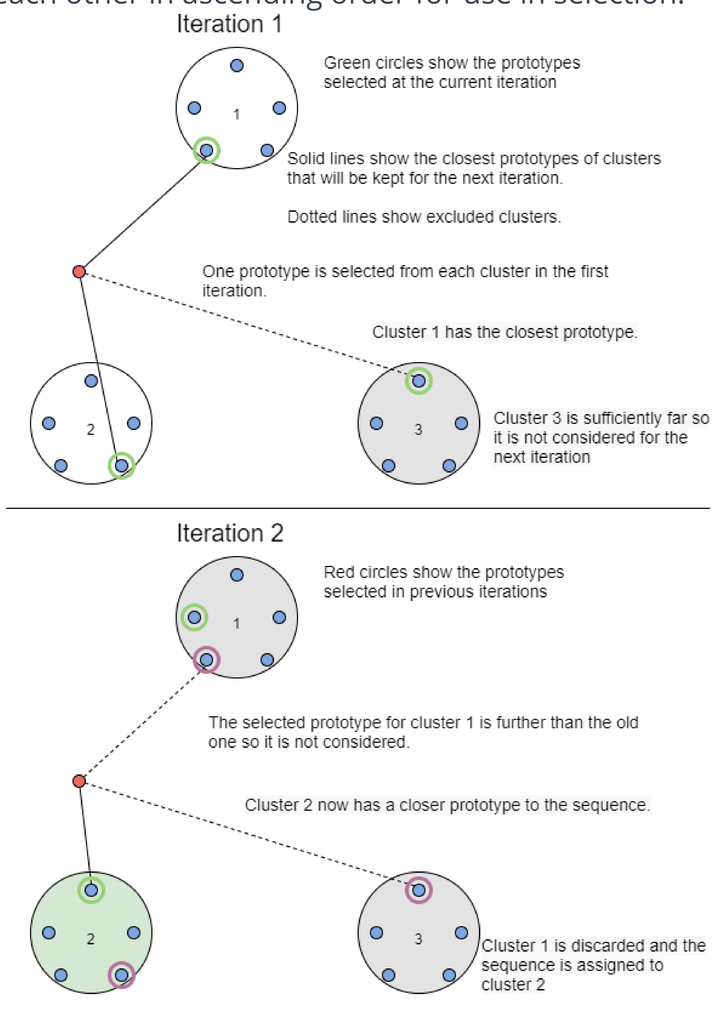
The posed research question is "For cluster assignment, which reference prototypes should be used to compute distance from?"

PROBLEM DESCRIPTION

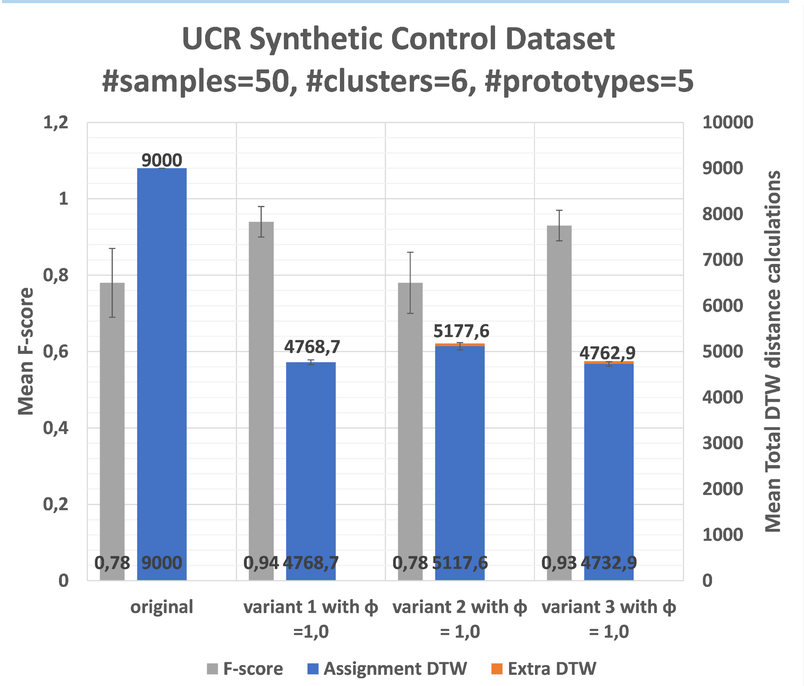
- Looking at the current process, we can see that 5 DTW distance calculations have to be done per cluster during the cluster assignment process. Calculating DTW distance is computationally expensive.
- An approach to optimize cluster assignment is to reduce the number of DTW distance calculations.
- Therefore, the goal is to use fewer prototypes for cluster assignment while continuing to represent a cluster with 5 prototypes.
- Assigning a sequence to the cluster with lowest mean distance to its prototypes can be problematic as mean is susceptible to outliers so an alternative needs to be explored that doesn't rely on the mean.

CONTRIBUTION

- A new iterative algorithm (with three variants) that searches for the closest prototype while excluding sufficiently far clusters.
- Parameter Φ controls how close a cluster should be.
- Variant 1 selects prototypes randomly at each iteration
- Variant 2 starts with a random prototype and selects the nearest one to the last selected one.
- Variant 3 orders prototypes by the average distance to each other in ascending order for use in selection.



RESULTS



CONCLUSION & FUTURE WORK

- The new algorithm shows a significant improvement in clustering speed across its three variants.
- Variants 1 & 3 show an improvement in clustering accuracy when compared to the original.
- Variant 2 fails to improve the clustering accuracy.
- Future work to be considered:
  - Optimize Φ.
  - Research other heuristics to select prototypes.
  - Test SeqClu and its improvement in their intended context (clustering network traffic) and research the ethical implications that the algorithm could have on the network users (e.g., clustering normal network traffic as malicious).