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 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 average distance between its prototypes and the • Variant 2 starts with a random prototype and selects the nearest one to the last selected one. • SeqClu uses Dynamic Time Warping (DTW) as distance • Variant 3 orders prototypes by the average distance to each other in ascending order for use in selection. Iteration 1 **RESEARCH QUESTION** Green circles show the prototypes selected at the current iteration \circ The posed research question is "For cluster assignment, which reference prototypes should be used to compute that will be kept for the next iteration. Dotted lines show excluded clusters One prototype is selected from each cluster in the first **PROBLEM DESCRIPTION** • Looking at the current process, we can see that 5 DTW (\mathbf{O}) distance calculations have to be done per cluster \circ \circ \circ 2 during the cluster assignment process. Calculating O/Iteration 2 • An approach to optimize cluster assignment is to Red circles show the prototypes selected in previous iterations reduce the number of DTW distance calculations. 0 \circ • Therefore, the goal is to use fewer prototypes for Q cluster assignment while continuing to represent a The selected prototype for cluster 1 is further than the old one so it is not considered. • Assigning a sequence to the cluster with lowest mean Cluster 2 now has a closer prototype to the sequence. distance to its prototypes can be problematic as mean is susceptible to outliers so an alternative needs to be 0 ${\rm O}$ \circ 0 \circ O \bigcirc

- 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 incoming sequence.
- metric instead of Euclidean distance.

distance from?"

- DTW distance is computationally expensive.
- cluster with 5 prototypes.
- explored that doesn't rely on the mean.

Solid lines show the closest prototypes of clusters

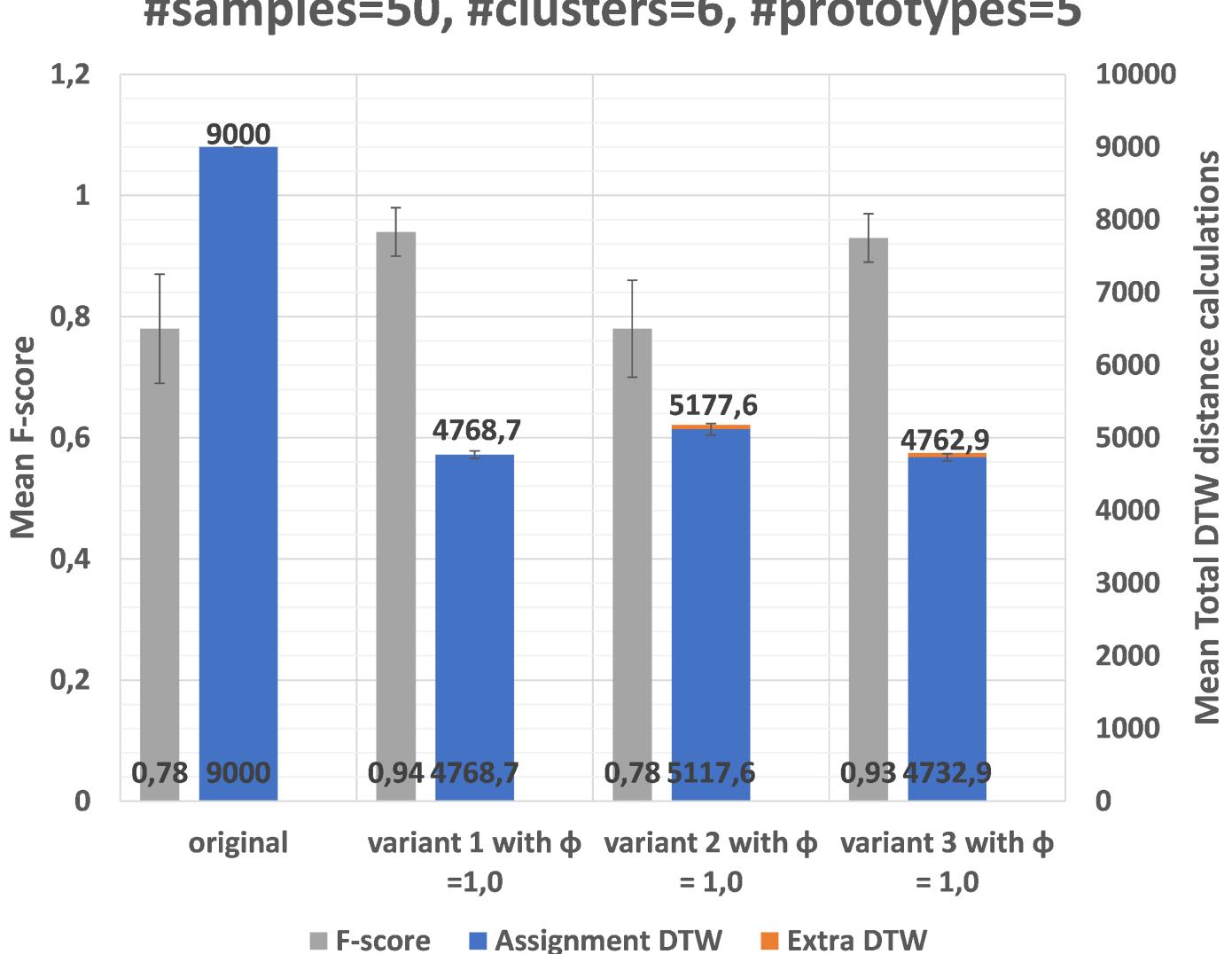
Cluster 1 has the closest prototype.



Cluster 3 is sufficiently far so it is not considered for the next iteration



Cluster 1 is discarded and the sequence is assigned to cluster 2



CONCLUSION & FUTURE WORK

- Future work to be considered:
- Optimize Φ.

- traffic as malicious).

RESULTS

UCR Synthetic Control Dataset #samples=50, #clusters=6, #prototypes=5

• 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.

 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