Cold start is coming: How to approximate the optimal set of prototypes for SeqClu

Author: Silviu Fucarev / Supervisor: Msc. Azqa Nadeem Responsible Professor: Dr. ir. S.E Verwer / Faculty of Electrical Engineering Mathematics and Computer Science

BACKGROUND INFORMATION

- SeqClu is an online sequence clustering algorithm
- Clusters are represented by 5 prototypes
- Sequences are assigned to the cluster with the smallest average DTW distance.
- The baseline initializes the clusters with 5 points from every cluster.

RESEARCH QUESTION

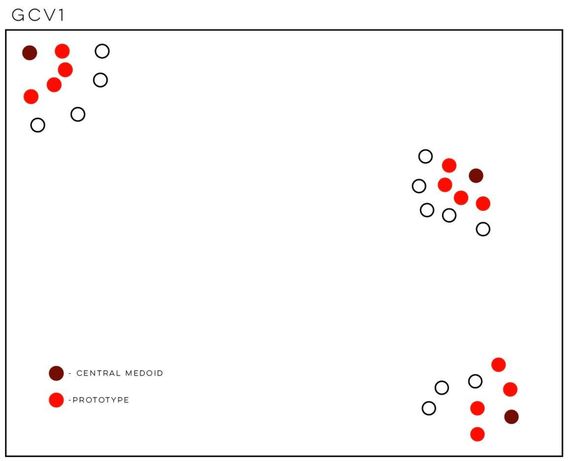
How to select the initial prototypes for SeqClu in an online setting?

PROCEDURE

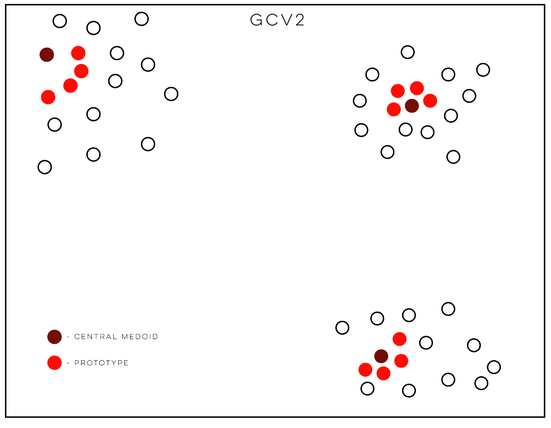
- Adapt well known heuristics to the online setting
- 2 Greedy Construction Heuristics: GCV1, GCV2
  - GCV1: Choose c furthest points and build clusters.
  - GCV2: Choose c furhtest points with tightest clusters.
- 2 Clustering Heuristics: K-Medoids++, K-Medoids Greedy
  - Offline K-Medoids on initial batch of points.
- Measure silhouette, accuracy (TP/N), average loss (distance to cluster)
- Estimate best initial batch size per initialization heuristic.

CONTACT INFORMATION

Name: Silviu Fucarev
Email: S.Fucarev@studnt.tudelft.nl
Supervised by: Azqa Nadeem
Responsible Professor: Sicco Verwer



GCV1 Greedily selects 3 points farthest away from each other as initial medoids, and their 4 closest neighbours as the other prototypes.



GCV2 picks the group of 5 points with the smallest sum of distances from the central medoid to the prototypes.
Next it picks the furthest cluster of 5 that also minimizes the distane between central medoid and prototypes until we have 3 clusters.

CONCLUSION

- GC Heuristics are preferred to clustering heuristics, and GCV1 is preferred to GCV2.
- GC achieves similar performance to best case scenario with a initial batch size of 25.
- GC Combined has potential to improve GCV1 but a suitable metric for decision making is not yet found.
- Clust. heuristics require batch sizes between 45-55 but fail with close clusters.
- K-Medoids++ is preferred to k-Medoids Greedy.
- Clust. heuristics are prone to cluster imbalance.