SeqClu-PV: An extension of online K-medoids to efficiently cluster sequences real-time

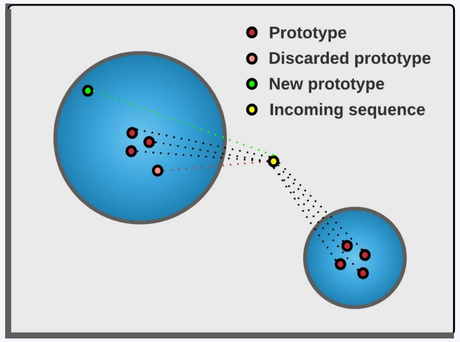
Ruben te Wierik, Azqa Nadeem, Dr.ir. Sicco Verwer

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

1. Problem

- *SeqClu* is an online variant of the *k*-medoids algorithm.
- Sequence data are clustered using *Dynamic Time Warping* (DTW), which is a computationally expensive distance measure.
- A cluster is represented with *p* most representative sequences, called prototypes, which requires little memory to maintain.
- The *SeqClu* algorithm computes the average distance between an incoming sequence and the prototypes of a cluster and assigns the sequence to the cluster that minimizes this average distance.
- The *SeqClu* algorithm is computationally expensive due to the many distance computations that need to be executed to assign an incoming sequence to a cluster.

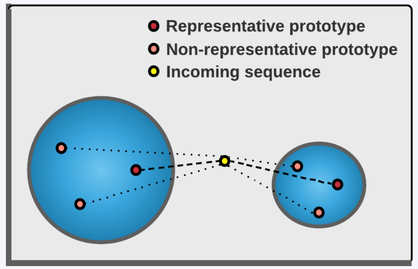


2. Research question

What combination of efficient decision making mechanisms for updating prototypes has the optimal balance between the number of computations required to cluster sequences in an online setting using the K-medoids algorithm and the cost incurred due to incorrect clustering?

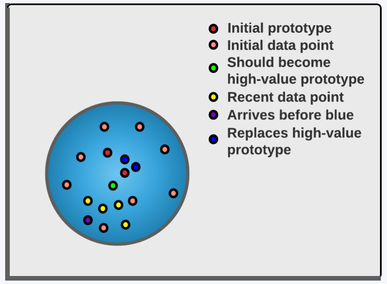
Practical algorithms for clustering streams of sequences in an online setting would be useful in a wide range of applications involving sequences of data that are generated real-time, such as the analysis of internet traffic, many Internet of Things (IoT) applications and real-time systems involving multimedia. This research focuses on when and how to update prototypes to identify improvements to existing online clustering algorithms.

3. Method

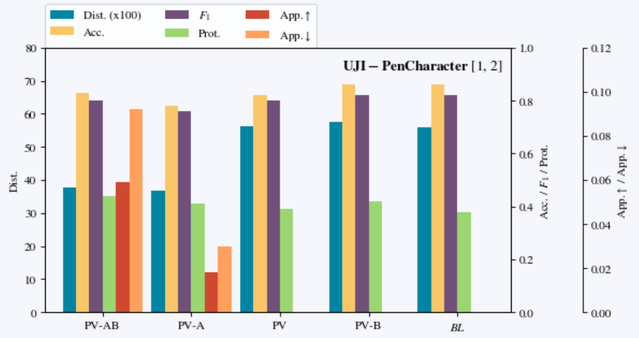


- Average distance to a cluster computed as the average distance to representative- instead of all prototypes.
- This idea could be the cause of sequences being assigned incorrectly, as is shown in the figure to the left.
- The problem is solved by using *representativeness.*

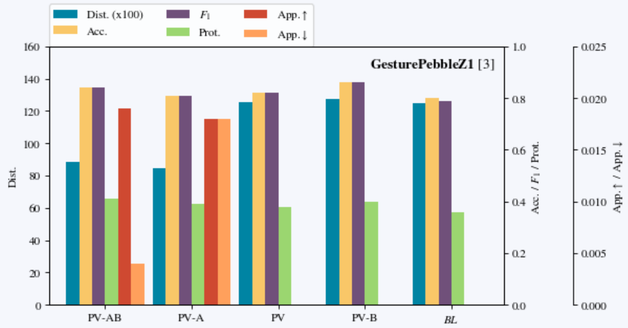
- Another problem with *SeqClu* is that it does not consider the similarity of prototypes to the earlier assigned sequences.
- The idea behind prototype voting is to keep track of how often an incoming sequence that is not a candidate prototype is closest to any of the prototypes.



4. Results



A: With approximation of distances | B: With buffering of candidate prototypes



A: With approximation of distances | B: With buffering of candidate prototypes

5. Conclusions

- Prototype voting is a powerful addition that results in a more optimal balance between computational cost and accuracy.
- *SeqClu-PV-A* variant reduces the number of distance computations
  - at the expense of accuracy for noisy and imbalanced data sets;
  - and increases the performance for well-separated and dense data sets.
- *SeqClu-PV-B* variant achieves an even more optimal balance between the number of distance computations and the cost incurred due to incorrect clustering at the expense of extra memory.

References

[1] Dheeru Dua and Casey Graff. {UCI} Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

[2] D. Llorens, F. Prat, A. Marzal, J. M. Vilar, M. J. Castro, J. C. Amengual, S. Barrachina, A. Castellanos, S. Espana, J. A. G ˜ omez, J. Gorbe, A. Gordo, V. Palaz ´ on, ´ G. Peris, R. Ramos Garijo, and F. Zamora. The UJIpenchars database: A pen-based database of isolated handwritten characters. Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, (January):2647–2651, 2008.

[3] Antigoni Mezari Mezari and Ilias Maglogiannnis. Gesture recognition using Symbolic Aggregate approximation and Dynamic Time Warping on Motion Data. ACM, 2018. doi: 10.1145/3154862.3154927.

Contact information

Ruben te Wierik

E-mail address: r.e.c.tewierik@student.tudelft.nl

Supervised by: Azqa Nadeem

Responsible professor: Sicco Verwer