Advanced Data Mining

Piotr Lipiński

List of Assignments 1 – time series clustering and classification

Assignment 1. *(4 points)*
a) Recall the definition of the DTW distance and write a function for evaluating it according to the algorithm described in our lecture.
b) Construct a few simple examples and verify the results of your function with some illustrations of the points that were matched.
c) Generate at least 10 random time series and for each of them create at least 100 random copies with various scaling and shifting.
d) Evaluate the DTW distance matrix between these time series using your function. If possible, increase the number of the time series and compare the computation time.

Assignment 2. *(4 points)*
Improve your DTW distance function in the following way:
- reduce the frequency of the time series or shrink the time series by removing some observations
- evaluate the DTW distance on the reduced time series
- transfer the assignment of the corresponding points from the reduced time series to the original time series
- evaluate the DTW distance on the original time series with the transferred assignment of the corresponding points
- try to reduce the distance by some local changes in the assignment (such as moving an assignment a few time instances back or forth)

Compare the results and the computation time of the improved function with the regular DTW function written in Assignment 1. You may also compare your approach with FastDTW ([1]).

Assignment 3. *(4 points)*
Please read the notebook *Introduction to Time Series Clustering* attached to the lecture notes.
a) Perform the clustering of daily water consumption profiles with DBA-k-means and compare it with the regular k-means approach.
b) Note that the daily water consumption profiles were aggregated into a type of day-of-week month-of-year mean profiles (7 x 12 = 84 time series). Please perform the clustering on the non-aggregated data (651 time series) also.
c) Please preprocess the input data with smoothing time series by a moving average (replacing the original value with an average of a few previous values) and repeat the clustering.
d) Find the Arrow Head Dataset and the Basic Motion Dataset and try to cluster them with the regular k-means as well as DBA-k-means (please focus only on time series clustering and ignore the class labels in these datasets).

Assignment 4. *(6 points)*
Please propose an extension of Random Forest algorithm for time series classification based on feature-based representation. You should propose a method to define the

efficient time windows in the time series for evaluating the features. Perform some experiments on the Arrow Head Dataset and the Basic Motion Dataset with the set of features including at least the mean, standard deviation and slope. Please compare your approach with Time Series Forest ([2]) and its implementations in the sktime package ([3]).

References:

[1] *S. Salvador, P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space". Intelligent Data Analysis, 5(11), 2007, pp.561-580.*

[2] *H. Deng, G. Runger, E. Tuv, M. Vladimir, "A time series forest for classification and feature extraction". Information Science 239, 2013, pp.142-153.*

[3] *M. Loning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines, F. J. Kiraly, "sktime: A Unified Interface for Machine Learning with Time Series". Workshop on Systems for ML at NeurIPS, 2019.*