CS 674 – Data Mining on Multimedia Data

HW 1
Due 2/22/17 at 4:30pm

A great deal of research in time series data mining is devoted to finding good representations and similarity measures, as they are the two key factors that greatly impact the efficiency and effectiveness of most time series data mining algorithms. In this homework, you will demonstrate your understanding and knowledge on these two fundamental topics in the mining of time series data, through the task of classification.

You will be given 5 time series datasets, and your job is to classify each dataset. One classic approach is to use 1-Nearest Neighbor (1-NN) classifier, with Euclidean Distance or Dynamic Time Warping as the distance measure on the raw data (i.e. no feature extraction or dimensionality reduction). It's been shown that 1-NN is very competitive, and it's one of the most widely used classifier for time series data. In this exercise, you will compare the classification accuracy of Euclidean Distance and Dynamic Time Warping. While there are many publicly available codes for this task, you must write your own code (including Euclidean distance, DTW, and 1-NN). For Dynamic Time Warping, there is a parameter w, the warping window size. Compare two versions of DTW: (1) no constraint on warping window size and (2) set warping window size to 20% of the length of the time series. The last part of the exercise is to try and improve the best accuracy. Some options to consider: use K-NN instead of 1-NN, learn the best warping window size for DTW, or something else.

You should have 4 accuracy results for each dataset: (1) Euclidean distance, (2) DTW with no constraint, (3) DTW with pre-defined warping window size (20%), and (4) your choice of improvement.

You can use any programming language of your choice. The only requirement is that you must write your own code and adhere to the honor code policy.

Format of the datasets: Each data file contains a M-by-N matrix, where M is the number of time series, and N is the length of time series + 1 (the first column of the matrix contains the class labels). Each dataset is split into training and test set for you. Report the classification accuracy on the test set.

To submit: source code, README, classification accuracy and a report (graphs showing the results, analysis, description of what you did, etc.)