MM811 - Multimedia Data Mining (Winter 2016)
Assignment 5

Due: March 21st, 2016 at 23:55pm

The goal of this assignment is to gain a better understanding of the topics covered in the course on feature selection and speeding up the process for finding similar time series.

For this assignment, you will be using the following datasets from UCR Time Series Archive (`http://www.cs.ucr.edu/~eamonn/time_series_data/`).

- *ECG5000* which is a set of 5000 time series each of length 140, broken down to 500 training set and 4500 test set.
- *Non-Invasive Fetal ECG Thorax1* which is a set of 1800+1965 time series each of length 750, broken down to 1800 training set and 1965 test set.

The datasets are available from the course eclass page. More information about the datasets and the file formats can be obtained from the UCR site.

The task is to find the distance between sequences in the test set to those in the training set, and for each sequence in the test set, use the label of its nearest neighbour (1-NN) to classify it. As the distance function, you will be implementing the Euclidean distance.

Task 1.
Write a program that finds the nearest neighbour for each test sequence, using a full Euclidean distance computation. (a) For each dataset, report the accuracy of the classification. (b) Measure the time for your distance calculation only (i.e. the time it takes to compute the Euclidean distance between a test sequence and a training sequence), ignoring the time for loading and any other computation; do this for all your test sequences. Report the mean and the standard deviation of the running time for each dataset.

Task 2.
You want to improve the efficiency of your algorithm for Task 1. For this, you will implement an early abandoning strategy where you will stop the distance computation

when you are certain the distance cannot be smaller than the current NN-distance. For each dataset, report accuracy, the mean and the standard deviation of the running time.

Task 3.
Compute and store DFT for each sequence; for your information, DFT is available in Matlab as a function named *fft*; you can also use public code from the web to compute the DFT as long as the source is cited. (a) Verify your DFT representation by computing the Euclidean distance for a pair of at least 5 DFT sequences and comparing them to the distances between respective time series. Report the sequences used and the result of your verification. (b) Perform Task 2 but this time on the DFT sequences. Do you see any difference in running time? Explain in either case. (c) Now you would be computing the distances only using 5, 10 and 20 first DFT coefficients. How does this affect the accuracy of the classifier and the running time? In each case and for each dataset, report accuracy, the mean and the standard deviation of the running time. (d) Did you take the symmetry property of DFT (as discussed in class) in your computation in Part b and c? Discuss the effect of taking and not taking it into account.

Deliverables A tarred and gzipped file submitted through eclass, which has the following pieces: (1) the program source code and a readme file describing how the code can be compiled or run, (2) a PDF report that discusses your experimental setup (such as the machine being used, its hardware and software spec) and your results and discussions for Tasks 1-3.