DBNT Data Assembly Readme

10/16/2013

Harrison Hunter

The data wrangler for the DBNT system is currently written in MATLAB.

The DBNT requires data as learning segment vectors in the following format:

[PatientID, SegmentID, n\*Params]

Where n refers to the number of distinct time slices within that segment. There will be 2+n\*len(Params) total items in each column.

Params (all for a given time slice) =

Mean The mean arterial pressure (map) for valid beats

Std The standard deviation of map over valid beats

Trend ...

Kurtosis ..

Skew .

Median Velocity

Median Acceleration

Ex.

3997097 0 6 10 10 7 10 10 8 …

The DBNT data wrangler writes to a .txt file in that format for each patient in the data source and for every available learning segment for each patient. The text file is named dynamically, defined below.

The top level script is dbnt\_dataasembly\_abp and it can be run by calling dbnt\_dataasembly\_abp(). This script loads the config file which has the following properties:

D.dataPath The path to the data

D.tmpDir The tmp directory to store in if requested

D.cachePath The cached directory to check if there is already processed data

D.segmentsFileName Where to save the segments as .mat if requested

The params for creating and segments and discretization:

D.overlapAmount

D.labelSeperators

D.cutOffClass

D.numberOfLagUnits

D.lagUnitLengthMinutes

D.timeSliceSize

D.runDistributed

The system currently has 8 files (including the config) of which 4 are helper functions or class definitions (BinarySearch, patientClass, map2classes, Mimic\_segment, HMM\_segment) and one is the config. Out of the three actual scripts here are their functionality:

dbnt\_dataasembly\_abp - This is the main file which loads the config, and holds the main logic for the data wrangling.

MIMIC\_data\_loader - This connects the matlab system to the data source.

Aggregation\_features - This defines the desired features and defines the methods that aggregate the data loaded into .mat files in the segments format defined by the HMM\_segment type.

Note the name of the .txt file is dynamically generated and is defined as:

data\_overlapAmount\_numberOfLagUnits\_lagUnitLengthMinutes\_timeSliceSize.txt

And will thus be different for different configurations, but will collide in the case where you use two different data sources, but the same parameters. This is an easy fix if desired, but we will standardize and assume it is run on the whole nfs directory.