THE PWM DATASETS

PIERRE-FRANÇOIS MARTEAU

UBS/VALORIA

JULY, 2007

1. Motivation

The artificial 'Pulse Width Modulation' (PWM) datasets have been defined to demonstrate a weakness in dynamic time warping (DTW) pseudo distance.

Basically, DTW does not penalize the matching of similar events having different time durations. For instance, the DTW similarity between any two time series chosen among the set given in Figure 3 below is the constant 0.

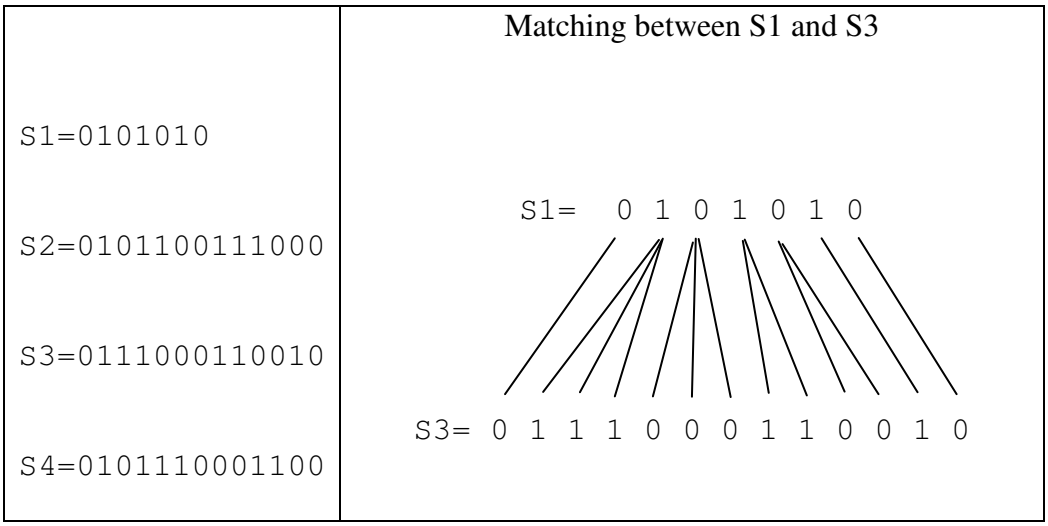

Fig. 1: Matching example between S1 and S3 time series

2. PWM Datasets

To demonstrate that this could affect recognition or classification of time series, the PWM datasets (PWM1 and PWM2) are defined having in mind a classification experiment with the same kind of conditions as those defined by E. Keogh & al. at UCR [1].

2.1 PWM1 dataset

The PWM1 dataset consists of artificial time series belonging to three categories. The considered task consists in classifying an input time series as one of three possible classes, '123' (a), '321' (*b*) or '132' (*c*). These classes are built as a sequence of three pulses having the same height but different widths: pulse '1' has the shortest width, while pulse '3' has the larger one and pulse '2' is in between.

To make it a little bit tougher, following Cylinder Bell Funnel artificial time series definition [2], random noise at four levels is added:

- a centered Gaussian noise with a standard deviation equal to 10% of the amplitude of each pulse is added,
- The onset time for each spike is selected uniformly inside a finite interval,
- The width of each spike is selected uniformly inside a finite interval that depends on its category ('1', '2' or '3').
- Time delays between two successive spikes are chosen uniformly inside a finite interval.

2.1.1 Time series definitions

Time series are generated according to their class as follows:

Class 123:

$$a(t) = \varepsilon(t) + \chi_{[s_1, w_1]}(t) + \chi_{[s_2, w_2]}(t) + \chi_{[s_3, w3]}(t)$$

Class 321:

$$a(t) = \varepsilon(t) + \chi_{[s_1, w_3]}(t) + \chi_{[s'_2, w_2]}(t) + \chi_{[s'_3, w_1]}(t)$$

Class 132:

$$a(t) = \varepsilon(t) + \chi_{[s_1, w_1]}(t) + \chi_{[s_2, w3]}(t) + \chi_{[s''_3, w_2]}(t)$$

where: $\varepsilon(t)$ is drawn from a standard normal distribution $N(0,1/10)$,

w_1 is an integer drawn uniformly from *[4, 8]*,

w_2 is an integer drawn uniformly from *[10, 16]*,

w_3 is an integer drawn uniformly from *[20, 28]*,

s_1, s'_1, s''_1 are integers drawn uniformly from *[2, 18]*,

s_2 is an integer drawn uniformly from $[s_1 + w_1, s_1 + s'_1 + w_1]$,

s_3 is an integer drawn uniformly from $[s_2 + w_2, s_2 + s''_1 + w_2]$,

s'_2 is an integer drawn uniformly from $[s_1 + w_3, s_1 + s'_1 + w_3]$,

s'_3 is an integer drawn uniformly from $[s'_2 + s''_1 + w_2, s'_2 + s''_1 + w_2]$,

s''_3 is an integer drawn uniformly from $[s_2 + w_3, s_2 + s''_1 + w_3]$,

$$\chi_{[\alpha,\beta]}(t) = \begin{cases} 0 & t < \alpha \\ 1 & \alpha \le t \le \beta \\ 0 & t > \beta \end{cases}$$

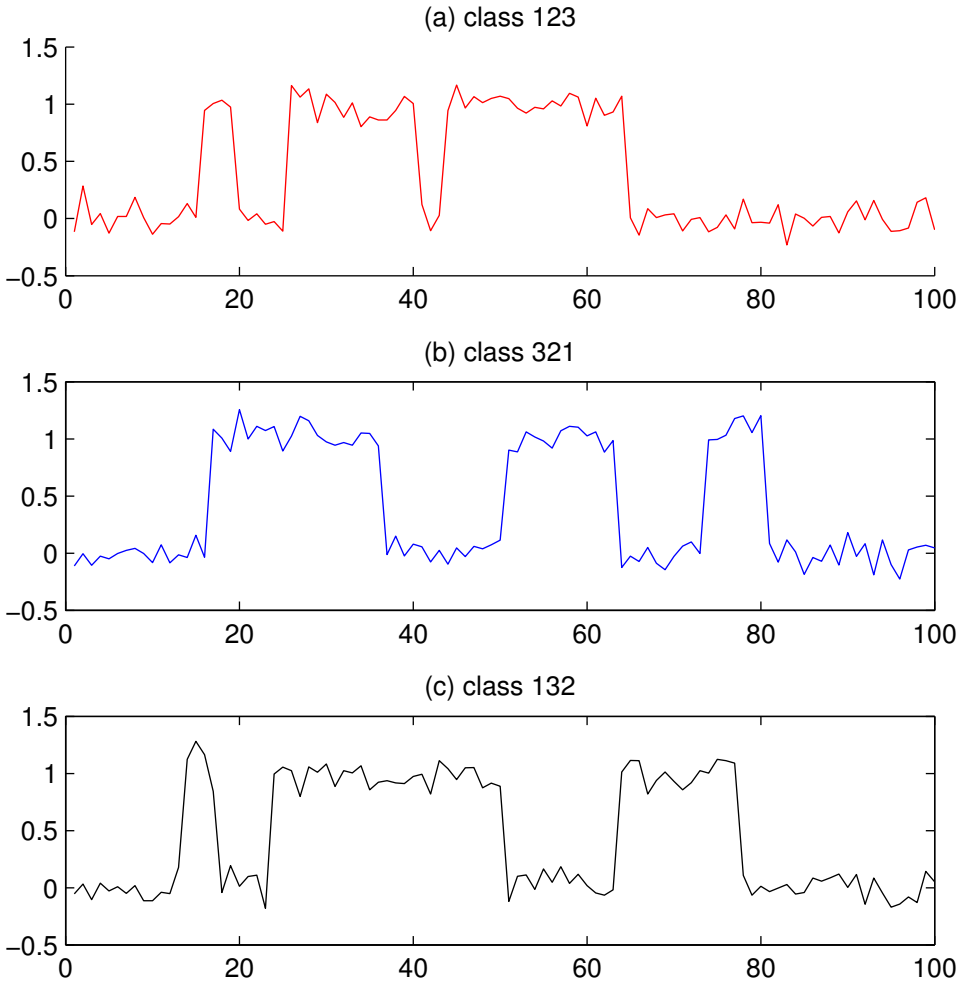

***Figure 2**: utterances for classes (a) 123 red, (b) 321 blue, (c) 132 black for the PWM1 dataset*

2.2 The PWM2 dataset

The PWM2 dataset consists of artificial time series belonging to three categories. The considered task consists in classifying an input time series as one of three possible classes, '123' (a), '321' (*b*) or '132' (*c*). These classes are built as a sequence of three subsequences of three pulses having the same height but different widths: subsequence '1' is composed of the shortest width pulses, while subsequence '3' is composed of the largest width pulses and subsequence '2' is composed of in between width pulses.

To make it a little bit tougher, following Cylinder Bell Funnel artificial time series definition [2], random noise at four levels is added:

- a centered Gaussian noise with a standard deviation equal to 10% of the amplitude of each pulse is added,
- The onset time for each pulse is selected uniformly inside a finite interval,
- The width of each pulse is selected uniformly inside a finite interval that depends on its category ('1', '2' or '3').
- Time delays between two successive pulses are chosen uniformly inside finite intervals that depend on the subsequence class.

2.1.1 Time series definitions

Time series are generated according to their class. Each time series contains three subsequences. Each subsequence has a fixed length of 25 samples. A subsequence is defined using three parameters: the starting time stamps t_0, the pulse width w and the pulse amplitude A. The subsequence $sq_{[t_0, w, A]}(t)$ is defined as follows:

$$sq_{[t_0,w,A]}(t) = \begin{cases} 0, & t < t_0 \\ A.(\chi_{[t_0,w]} + \chi_{[t_1,w]} + \chi_{[t_2,w]}), & t_0 \le t \le t_0 + 25 \\ 0, & t > t_0 + 25 \end{cases}$$

where:

t_1 is an integer drawn uniformly from $[t_0+w+1,\ t_0+w+25-3.w-1]$

t_2 is an integer drawn uniformly from $[t_1+w+1,\ t_1+w+25-2.w]$

and $\chi_{[\alpha,\beta]}(t) = \begin{cases} 0 & t < \alpha \\ 1 & \alpha \le t \le \beta \\ 0 & t > \beta \end{cases}$

Finally the three classes are defined as follows:

Class 123:

$$a(t) = -1 + \varepsilon(t) + sq_{[t_0,w_1,2]} + sq_{[t_0+25,w_2,2]} + sq_{[t_0+50,w_3,2]}$$

Class 321:

$$a(t) = -1 + \varepsilon(t) + sq_{[t_0,w_3,2]} + sq_{[t_0+25,w_2,2]} + sq_{[t_0+50,w_1,2]}$$

Class 132:

$$a(t) = -1 + \varepsilon(t) + sq_{[t_0,w_1,2]} + sq_{[t_0+25,w_3,2]} + sq_{[t_0+50,w_2,2]}$$

where: $\varepsilon(t)$ is drawn from a standard normal distribution $N(0,1/10)$,

t_0 is an integer drawn uniformly from *[2, 32]*,

w_1 is an integer drawn uniformly from *[2, 3]*,

w_2 is an integer drawn uniformly from *[4, 5]*,

w_3 is an integer drawn uniformly from *[6, 7]*

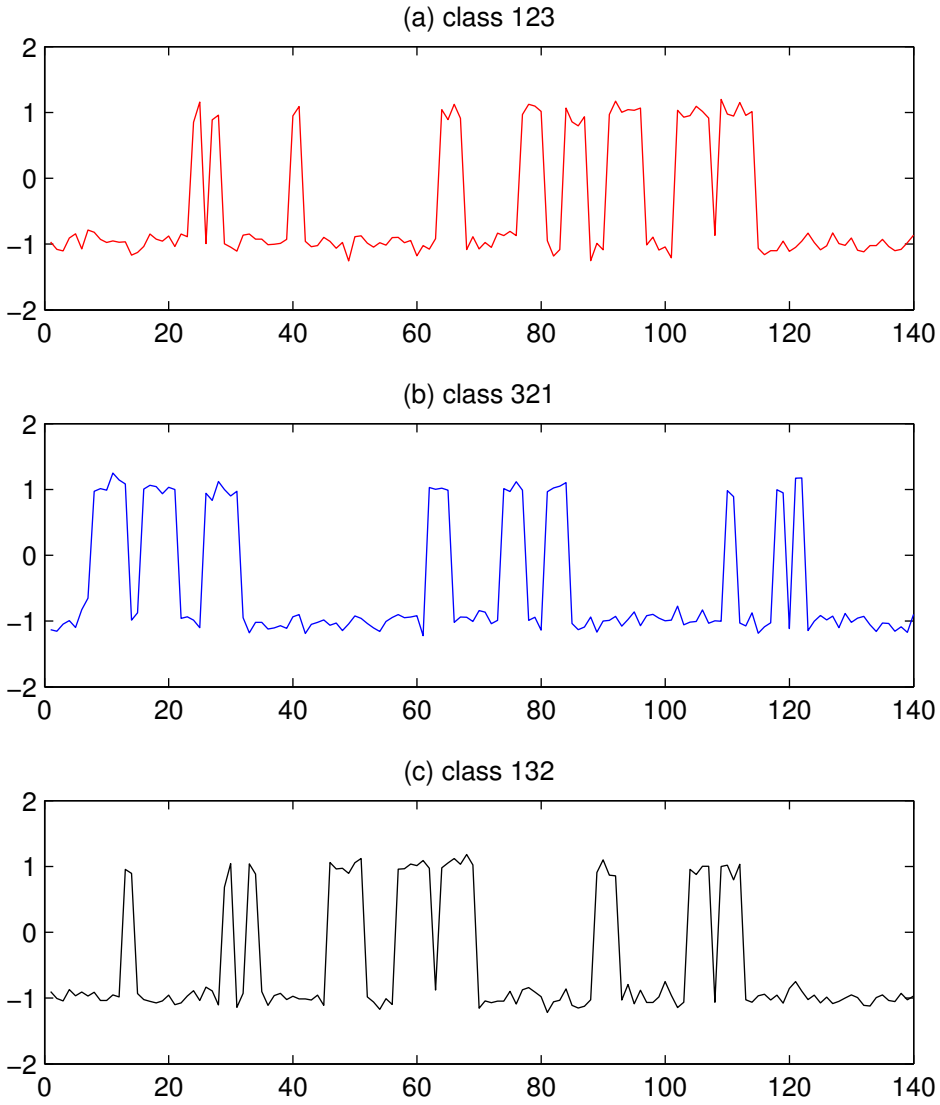

***Figure 3**: utterances for classes (a) 123 red, (b) 321 blue, (c) 132 black for the PWM2 dataset*

3. The files

The 3 classes are equi-likely inside the train and test datasets for PWM1 and PWM2:

TEST_PWM1 contains 600 time series, 200 per classes

TRAIN_PWM1 contains 60 time series, 20 per classes.

TEST_PWM2 contains 300 time series, 100 per classes

TRAIN_PWM2 contains 30 time series, 10 per classes.

File Format

TEST_PWM1, TEST_PWM2, TRAIN_PWM1 and TRAIN_PWM2 are ASCII files in which each line begins with a label (a number) that characterizes the class (1 for class 123, 2 for class 321 and 3 for class 132; the time series (1D data) samples are then coded in the rest of the line as a sequence of floating point number in ASCII format.

PWM1 and PWM2 datasets can be downloaded at the following URL:

http://www-valoria.univ-ubs.fr/Pierre-Francois.Marteau/PWM

References

[1] Keogh, E., Xi, X., Wei, L. & Ratanamahatana, C. A. (2006). The UCR Time Series Classification/Clustering Homepage: http://www.cs.ucr.edu/~eamonn/time_series_data/

[2] N., Saito. Local feature extraction and its application using a library of bases. PhD thesis, Yale University, 1994.