Scaling up Dynamic Time Warping for Datamining Applications

Eamonn J. Keogh Michael J. Pazzani

Department of Information and
Computer Science
University of California
Irvine, California 92697 USA
Phone (949) 824-7210
{eamonn,pazzani}@ics.uci.edu

ABSTRACT

There has been much recent interest in adapting data mining algorithms to time series databases. Most of these algorithms need to compare time series. Typically some variation of Euclidean distance is used. However, as we demonstrate in this paper, Euclidean distance can be an extremely brittle distance measure. Dynamic time warping (DTW) has been suggested as a technique to allow more robust distance calculations, however it is computationally expensive. In this paper we introduce a modification of DTW which operates on a higher level abstraction of the data, in particular, a Piecewise Aggregate Approximation (PAA). Our approach allows us to outperform DTW by one to two orders of magnitude, with no loss of accuracy.

Keywords

Time series, similarity measures, Dynamic Time Warping.

1. INTRODUCTION

Time series are a ubiquitous form of data occurring in virtually every scientific discipline and business application. There has been much recent work on adapting data mining algorithms to time series databases. For example, Das et al attempt to show how association rules can be learned from time series [5]. Debregeas and Hebrail [6] demonstrate a technique for scaling up time series clustering algorithms to massive datasets. Keogh and Pazzani introduced a new, scaleable time series classification algorithm [12]. Almost all algorithms that operate on time series data need to compute the similarity between time series. Euclidean distance, or some extension or modification thereof, is typically used. However, Euclidean distance can be an extremely brittle distance measure. Consider the clustering produced by Euclidean distance in Figure 1. Sequence 3 is judged as most similar to the line in sequence 4, yet it appears more similar to 1 or 2.

KDD 2000, Boston, MA USA

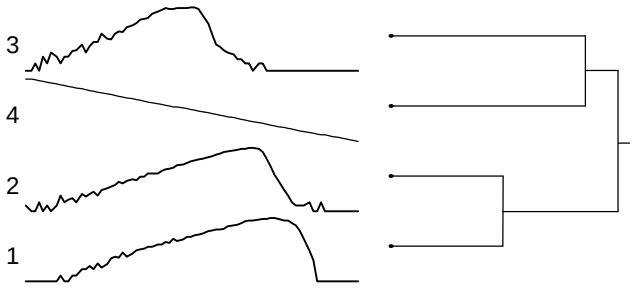

Figure 1. An unintuitive clustering produced by the Euclidean distance measure. Sequences 1 to 3 are astronomical time series [7]. Sequence 4 is simply a straight line with the same mean and variance as the other sequences.

The reason why Euclidean distance may fail to produce an intuitively correct measure of similarity between two sequences is because it is very sensitive to small distortions in the time axis. Consider Figure 2.A. The two sequences have approximately the same overall shape, but the shapes are not aligned in the time axis. The nonlinear alignment shown in Fig 2.B would allow a more sophisticated distance measure to be calculated.

A method for achieving such alignments has long been known in the speech processing community [20]. The technique, Dynamic Time Warping (DTW), was introduced to the data mining community by Berndt and Clifford [3]. Although they demonstrate the utility of the approach, they acknowledge that the algorithms time complexity is a problem and that "…performance on very large databases may be a limitation".

As an example of the utility of DTW compare the clustering shown in Figure 1 with Figure 3.

In this paper we introduce a technique which speeds up DTW by a large constant. The value of the constant is data dependent but is typically one to two orders of magnitude. The algorithm, Piecewise Dynamic Time Warping (PDTW), takes advantage of the fact that we can efficiently approximate most time series by a piecewise aggregate approximation.

The rest of this paper is organized as follows. Section 2 contains a review of the classic DTW algorithm. Section 3 introduces the Piecewise Aggregate Approximation and PDTW algorithm. In Section 4 we experimentally compare DTW, PDTW and Euclidean distance on several real world datasets. Section 5 contains a discussion of related work. Section 6 contains our conclusions.

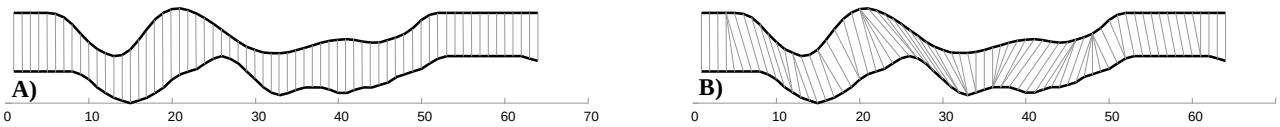

Figure 2. Two sequences from an Australian Sign Language dataset. Note that while the sequences have an overall similar shape, they are not aligned in the time axis. Euclidean distance, which assumes the ith point on one sequence is aligned with ith point on the other (A), will produce a pessimistic dissimilarity measure. A nonlinear alignment (B) allows a more sophisticated distance measure to be calculated.

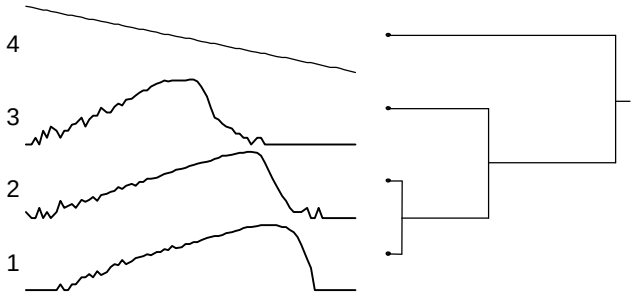

Figure 3. When the dataset used in Figure. 1 is clustered using DTW the results are much more intuitive.

2. DYNAMIC TIME WARPING

Suppose we have two time series Q and C, of length n and m respectively, where:

$$Q = q_1,q_2,\ldots,q_i,\ldots,q_n \quad (1)$$

$$C = c_1,c_2,\ldots,c_j,\ldots,c_m \quad (2)$$

To align two sequences using DTW we construct an n-by-m matrix where the (i^{th}, j^{th}) element of the matrix contains the distance $d(q_i,c_j)$ between the two points q_i and c_j (With Euclidean distance, $d(q_i,c_j) = (q_i - c_j)^2$). Each matrix element ($i$,$j$) corresponds to the alignment between the points q_i and c_j. This is illustrated in Figure 4. A warping path W, is a contiguous (in the sense stated below) set of matrix elements that defines a mapping between Q and C. The k^{th} element of W is defined as $w_k = (i,j)_k$ so we have:

$$W = w_1, w_2, \ldots,w_k,\ldots,w_K \quad \max(m,n) \le K < m+n-1 \quad (3)$$

The warping path is typically subject to several constraints.

- **Boundary conditions:** $w_1 = (1,1)$ and $w_K = (m,n)$, simply stated, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.
- **Continuity:** Given $w_k = (a,b)$ then $w_{k-1} = (a',b')$ where $a–a' \le 1$ and $b-b' \le 1$. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).
- **Monotonicity:** Given $w_k = (a,b)$ then $w_{k-1} = (a',b')$ where $a–a' \ge 0$ and $b-b' \ge 0$. This forces the points in W to be monotonically spaced in time.

There are exponentially many warping paths that satisfy the above conditions, however we are interested only in the path which minimizes the warping cost:

$$DTW(Q,C) = \min\left\{ \sqrt{\sum_{k=1}^{K} w_k} \Big/ K \right\} \quad (4)$$

The K in the denominator is used to compensate for the fact that warping paths may have different lengths.

This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i,j)$ as the distance $d(i,j)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements:

$$\gamma(i,j) = d(q_i,c_j) + \min\{ \gamma(i-1,j-1) , \gamma(i-1,j) , \gamma(i,j-1) \} \quad (5)$$

The Euclidean distance between two sequences can be seen as a special case of DTW where the k^{th} element of W is constrained such that $w_k = (i,j)_k$, i = j = k. Note that it is only defined in the special case where the two sequences have the same length. The time complexity of DTW is O(nm).

This review of DTW is necessarily brief; we refer the interested reader to [16] for a more detailed treatment.

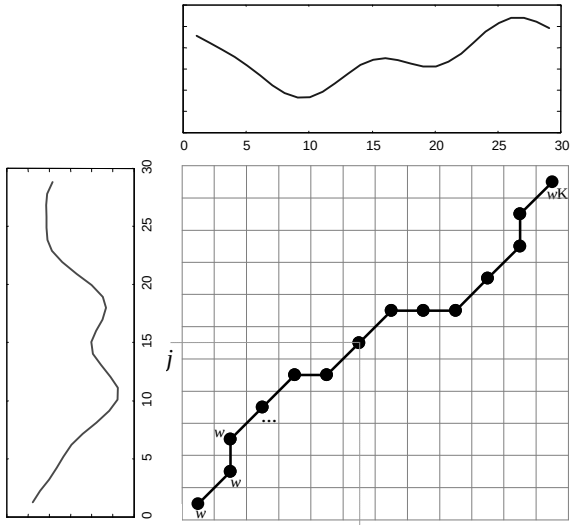

Figure 4. An example warping path.

3. A HIGHER LEVEL REPRESENTATION

In this section we introduce the piecewise aggregate approximation and a DTW algorithm for the representation.

3.1 The piecewise aggregate representation

We denote a time series query as $X = x_1,\ldots,x_n,$. Let N be the dimensionality of the transformed time series we wish to work with ($1 \le N \le n$). For convenience, we assume that N is a factor of n. This is not a requirement of our approach, however it does simplify notation.

A time series X of length n is represented in N space by a vector $\overline{X} = \overline{x}_1,\ldots,\overline{x}_N$. The ith element of $\overline{X}$ is calculated by the following equation:

$$\overline{x}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \quad (6)$$

Simply stated, to reduce the data from n dimensions to N dimensions, the data is divided into N equi-sized "frames". The mean value of the data falling within a frame is calculated and a vector of these values becomes the data reduced representation. Figure 5 illustrates this notation. The complicated subscripting in Eq. 6 is just to insure that the original sequence is divided into the correct number and size of frames.

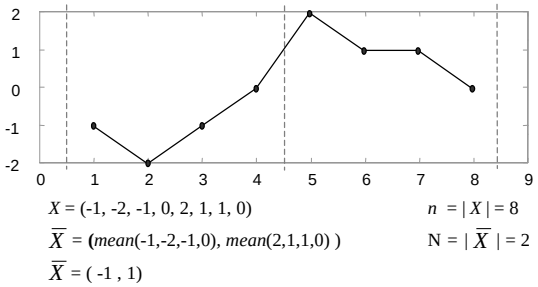

Figure. 5: An illustration of the data reduction technique utilized in this paper. A time series consisting of eight (*n*) points is projected into two (*N*) dimensions. The time series is divided into two (*N*) frames and the mean of each frame is calculated. A vector of these means becomes the data reduced representation.

Two special cases worth noting are when $N = n$ the transformed representation is identical to the original representation. When $N = 1$ the transformed representation is simply the mean of the original sequence. More generally the transformation produces a piecewise constant approximation of the original sequence, we therefore call our approach Piecewise Aggregate Approximation (PAA). Figure 6 illustrates a natural time series and its PAA approximation.

We denote the ratio of the length of the original time series to the length of its PAA representation, the compression rate c.

$$c = n/N \tag{7}$$

In choosing a value for c there is a classic tradeoff between memory savings and fidelity. In this work we do not address the problem of choosing the "best" compression rate. The "best" compression rate depends on the structure of the data itself and the task at hand (i.e. clustering/classification/retrieval etc). For most applications the best approach may be to have an expert interact with the data and choose this parameter, although automated approaches to similar problems have been suggested [22,15].

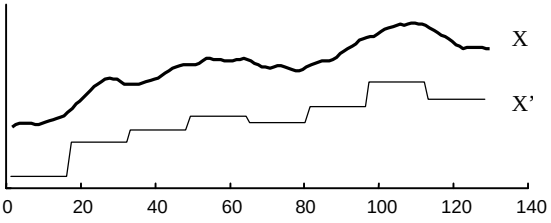

Figure 6: The sequence X **and its Piecewise Aggregate Approximation** X'.

3.2 Warping with the PAA representation

In Section 2 we showed how to perform dynamic time warping on two sequences Q and C. Here we will show how to perform dynamic time warping using the reduced dimensionality versions of Q and C, which we denote $\overline{Q}_i$ and $\overline{C}_i$ respectively. For clarity we call the algorithm defined on the reduced dimensionality representation Piecewise Dynamic Time Warping (PDTW).

To align two sequences using PDTW we construct an N-by-M matrix where the (i^{th}, j^{th}) element of the matrix contains the distance $d(\overline{Q}_i, \overline{C}_i)$ between the two elements $\overline{Q}_i$ and $\overline{C}_i$. The distance between two elements is defined as the square of the distance between them:

$$d(\overline{Q}_i, \overline{C}_i) = (\overline{Q}_i - \overline{C}_i)^2 \tag{8}$$

Apart from this modification the matrix-searching algorithm is essentially unaltered. Equation 5 is modified to reflect the new distance measure:

$$\gamma(i,j) = d(\overline{Q}_i, \overline{C}_i) + \min\{ \gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1) \} \tag{9}$$

When reporting the DTW distance between two time series (Eq. 4) we compensated for different length paths by dividing by K, the length of the warping path. We need to do something similar for PDTW but we cannot use K directly, because elements in the warping matrix now correspond to aggregate segments of data and we would like PDTW to be measured in the same units as DTW to facilitate comparison between the two measures. To compensate for this we can use a distance measure that is similar to Eq. 4 but where the denominator is the square root of the compression rate.

$$PDTW(\overline{Q}, \overline{C}) = \min\left\{ \sqrt{\sum_{k=1}^{K} w_k} \Big/ \sqrt{c} \right. \tag{10}$$

Because the length of the warping path is measured in the same units as DTW we have:

$$\text{PDTW}(\overline{Q}, \overline{C}) \cong \text{DTW}(Q,C) \tag{11}$$

Figure 7 shows strong visual evidence that SDTW finds alignments that are very similar to those produced by DTW. In the next section we will provide strong experiment evidence to the same effect.

The time complexity for a PDTW is O(NM), where $M = m/c$ and $N = n/c$. The time complexity for the original DTW algorithm is O(nm). So the speedup obtained by PDTW should be O(nm)/O(MN) which is O(c^2).

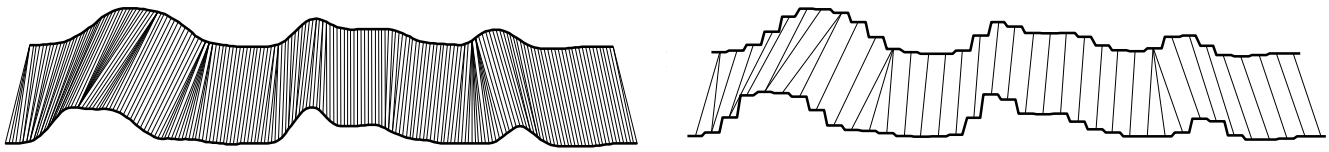

Figure 7: A) Two similar time series and the alignment between them, as discovered by DTW. B) The same time series in their PAA representation, and the alignment discovered by PDTW. This presents strong visual evidence that PDTW finds approximately the same warping as DTW.

4. EXPERIMENT RESULTS

We are interested in two properties of the proposed approach. The speedup obtained over the classic DTW algorithm and the quality of the alignment. In general, the quality of the alignment is subjective, so we designed experiments that indirectly, but objectively measure it.

4.1 Clustering

For our clustering experiments we utilized two datasets, one natural and one synthetic.

1) The Australian Sign Language (ASL) dataset from the UCI KDD archive [2]. The dataset consists of various sensors that measure the X-axis position of a subject's right hand while signing one of 95 words in Australian Sign Language.

2) The Cylinder-Bell-Funnel (CBF) synthetic dataset as used in [11,17,19]. This dataset contains three classes, which are generated by the following equations.

$$c(t) = (6+\eta) \bullet X_{[a,b]}(t) + \varepsilon(t)$$
$$b(t) = (6+\eta) \bullet X_{[a,b]}(t) \bullet (t-a)/(b-a) + \varepsilon(t)$$
$$f(t) = (6+\eta) \bullet X_{[a,b]}(t) \bullet (b-a)/(b-t) + \varepsilon(t)$$

$$X_{[a,b]} = \begin{cases} 0 & t < a \\ 1 & a \le t \le b \\ 0 & t > b \end{cases}$$

Where η and $\varepsilon(t)$ are drawn from a standard normal distribution N(0,1), a is an integer drawn uniformly from the range [16,32] and (b-a) is an integer drawn uniformly from the range [32, 96].

Figure 8 shows some examples of the Cylinder and Funnel class (members of the Bell class look like mirror images of the Funnel class).

For every possible pairing of the ten words in the ASL dataset, we clustered the 10 corresponding sequences, using group-average hierarchical clustering. At the lowest level of the corresponding dendrogram, the clustering is subjective. However, the highest level of the dendrogram (i.e. the first bifurcation) should divide the data into the two classes. Any dendrogram that correctly partitions the data in this fashion we consider correct and any other partition we consider incorrect. There are 34,459,425 possible ways to cluster 10 items, of which 11,025 of them correctly separate the two classes, so the default rate for an algorithm which guesses randomly is only 0.031%.

We performed the same experiments for the CBF dataset, with every possible pairing of the three classes. Figure 8 shows the results of one experiment with the Cylinder and Funnel classes. Here we had the luxury of unlimited data so we ran each experiment 100 times and averaged the results.

We compared four distance measures:

1) **DTW**: The classic dynamic time warping algorithm as presented in Section 2.

2) **PDTW**: The piecewise dynamic time warping algorithm proposed in this paper.

3) **Euclidean**: We also tested Euclidean distance measure to facilitate comparison to the large body of literature that utilizes this distance measure.

4) **PEuclidean**: Because it might argued that any increased accuracy of PDTW was due solely to the smoothing effects of the piecewise aggregate approximation, we also tested the Euclidean measure using the PAA representation.

Table 1 summarizes the results.

Distance Measure	ASL		CBL	
	Mean Time (Seconds)	Correct Clusterings (percentage)	Mean Time (Seconds)	Correct Clusterings (percentage)
DTW	174.4	48.8	519.2	92.1
PDTW	3.7	51.1	24.1	93.3
Euclidean	2.1	4.4	0.49	3.2
PEuclidean	2.3	4.4	0.62	4.8

Although the Euclidean distance can be quickly calculated, it performance is only a little better than random. While the smoothing effect of the PAA representation does help slightly for the CBL dataset, both of the Euclidean based metrics have great difficulty differentiating between two classes in both datasets. Both DTW and PDTW have essentially the same high accuracy, but PDTW faster by a factor of 47 for the ASL dataset and a factor of 21.5 for the CBL dataset.

5. RELATED WORK

Dynamic time warping has enjoyed success in many areas where its time complexity is not an issue. It has been used in gesture recognition [9], robotics [21], speech processing [18], manufacturing [10] and medicine [4].

Conventional DTW, however, is much too slow for searching large databases. For this problem, Euclidean distance, combined with an indexing scheme is typically used. Faloutsos et al, extract the first few Fourier coefficients from the time series and use these to project the data into multi-dimensional space [8]. The data can then be indexed with a multi-dimensional indexing structure such as a R-tree. Keogh and Pazzani address the problem by de-clustering the data into bins, and optimizing the data within the bins to reduce search times [12].

6. CONCLUSIONS

The most important contribution of this paper is to show that to Euclidean distance metric, although popular, is an extremely brittle distance measure that degrades rapidly in the presence of time axis distortion. We reintroduced DTW to the KDD community and demonstrated a modification of DTW that exploits a higher level representation of time series data to produce one to two orders of magnitude speed-up with no decrease in accuracy. We experimentally demonstrated our approach on several real world datasets and showed a speedup of one to two orders of magnitude.

REFERENCES

[1] Agrawal, R., Lin, K. I., Sawhney, H. S., & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in times-series databases. In VLDB, September.

[2] Bay, S. (1999). UCI Repository of Kdd databases [http://kdd.ics.uci.edu/]. Irvine, CA: University of California, Department of Information and Computer Science

[3] Berndt, D. & Clifford, J. (1994) Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on Knowledge Discovery in Databases*. Seattle, Washington.

[4] Caiani, E.G., Porta, A., Baselli, G., Turiel, M., Muzzupappa, S., Pieruzzi, F., Crema, C., Malliani, A. & Cerutti, S. (1998) Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume. *IEEE Computers in Cardiology*. Vol. 25 Cat.

[5] Das, G., Lin, K., Mannila, H., Renganathan, G. & Smyth, P. (1998). Rule discovery form time series. *Proc. of the 4th International Conference of Knowledge Discovery and Data Mining*. pp 16-22, AAAI Press.

[6] Debregeas, A. & Hebrail, G. (1998). Interactive interpretation of Kohonen maps applied to curves. *Proc. of the 4th International Conference of Knowledge Discovery and Data Mining*. pp 179-183, AAAI Press.

[7] Derriere, S. (1998) D.E.N.I.S strip 3792: [http://cdsweb.u-strasbg.fr/DENIS/qual_gif/cpl3792.dat]

[8] Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. *In Proc. ACM SIGMOD Conf*., Minneapolis, May.

[9] Gavrila, D. M. & Davis,L. S.(1995). Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. *In International Workshop on Automatic Face- and Gesture-Recognition*. IEEE Computer Society.

[10] Gollmer, K., & Posten, C. (1995) Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses. *On-Line Fault Detection and Supervision in Chemical Process Industries.*

[11] Kadous, M. W. (1999) Learning comprehensible descriptions of multivariate time series. *In Proc. of the 16th International Machine Learning Conference*. Morgan Kaufmann.

[12] Keogh, E., & Pazzani, M. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *Proc. of the 4th International Conference of Knowledge Discovery and Data Mining*. pp 239-241, AAAI Press.

[13] Keogh, E., & Pazzani, M. (1999). An indexing scheme for fast similarity search in large time series databases. *In Proc. of the 11th International Conference on Scientific and Statistical Database Management*.

[14] Keogh, E., & Pazzani, M. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases. *In 4th Pacific- Asia Conference on Knowledge Discovery and Data Mining*. Kyoto, Japan

[15] Keogh, E., Smyth, P. (1997). A probabilistic approach to fast pattern matching in time series databases. *Proc. of the 3rd International Conference of Knowledge Discovery and Data Mining*. pp 24-20, AAAI Press.

[16] Kruskall, J. B. & Liberman, M. (1983). The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules. Addison-Wesley.

[17] Manganaris, S. (1997). Supervised classification with temporal data. PhD thesis, Computer Science Department, School of Engineering, Vanderbilt University

[18] Rabiner, L. & Juang, B. (1993). Fundamentals of speech recognition. Englewood Cliffs, N.J, Prentice Hall.

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

[20] Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26.

[21] Schmill, M., Oates, T. & Cohen, P. (1999). Learned models for continuous planning. *In Seventh International Workshop on Artificial Intelligence and Statistics*.

[22] Shatkay, H., & Zdonik, S. (1996). Approximate queries and representations for large data sequences. Proc. 12th IEEE International Conference on Data Engineering. pp 546-553.

[23] Yi, B. K., Jagadish, H. V., Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. *In Proc. of the 14th International Conference on Data Engineering*. pp 201-208.

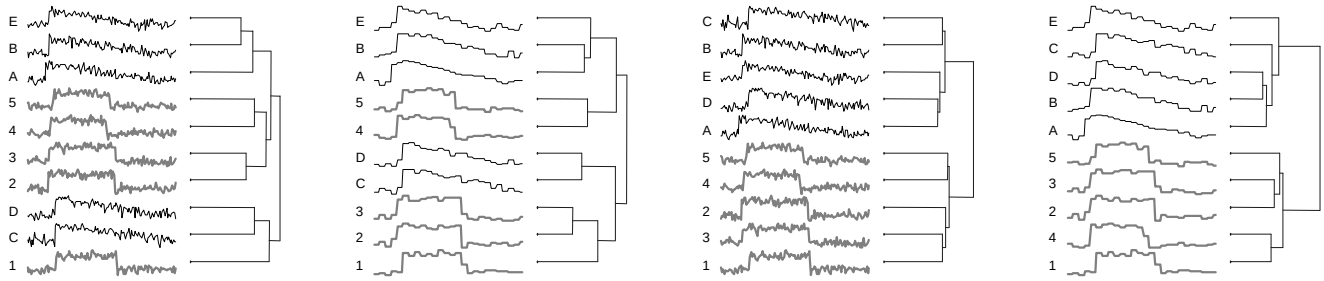

Figure 8. An example of a single clustering experiment on the cylinder-bell-funnel dataset. The time series 1 to 5 are members of the cylinder class. The time series A to E are members of the funnel class. The Euclidean distance metric has difficulty in differentiating between the two classes, but both DTW and PDTW correctly separate the two with the first bifurcation of the dendrogram