Inaccuracies of Shape Averaging Method Using Dynamic Time Warping for Time Series Data

Vit Niennattrakul and Chotirat Ann Ratanamahatana

Department of Computer Engineering, Chulalongkorn University
Phayathai Rd., Pathumwan, Bangkok 10330 Thailand
{g49vnn, ann}@cp.eng.chula.ac.th

Abstract. Shape averaging or signal averaging of time series data is one of the prevalent subroutines in data mining tasks, where Dynamic Time Warping distance measure (DTW) is known to work exceptionally well with these time series data, and has long been demonstrated in various data mining tasks involving shape similarity among various domains. Therefore, DTW has been used to find the *average* shape of two time series according to the optimal mapping between them. Several methods have been proposed, some of which require the number of time series being averaged to be a power of two. In this work, we will demonstrate that these proposed methods cannot produce the *real* average of the time series. We conclude with a suggestion of a method to potentially find the shape-based time series average.

Keywords: Time Series, Shape Averaging, Dynamic Time Warping.

1 Introduction

The need to find the template or the data representative from a group of time series is prevalent in major data mining tasks' subroutines [2][6][7][9][10][14][16][19]. These include query refinement in Relevance Feedback [16], finding the cluster centers in *k*-means clustering algorithm, and template calculation in speech processing or pattern recognition. Various algorithms have been applied to calculate these data representations, often times we simply call it a data average. A simple averaging technique uses Euclidean distance metric. However, its one-to-one mapping nature is unable to capture the actual average shape of the two time series. In this case, shape averaging algorithm, Dynamic Time Warping, is much more appropriate [8].

In shape-based time series data, shape averaging method should be considered. However, most work involving time series averaging appear to avoid using DTW in spite of its dire need in the shape-similarity-based calculation [2][5][6][7][9][10] [13][14][19] without providing sufficient reasons other than simplicity. For those who use *k*-means clustering, Euclidean distance metric is often used for time series average. This is also true in other domains such as speech recognition and pattern recognition [1][6][9][14], which perhaps is a good indicator flagging problems in DTW averaging method.

Y. Shi et al. (Eds.): ICCS 2007, Part I, LNCS 4487, pp. 513 – 520, 2007.

Despite many proposed shape averaging algorithms, most of them provide method for specific domains [3][11][12], such as evoked potential in medical domains. In particular, after surveying related publications in the past decade, there appears to be only one proposed by Gupta et al. [8], who introduced the shape averaging using Dynamic Time Warping, and has been the basis for all subsequent work involving shape averaging. As shown in Figure 1 (a), the average is done in pairs, and the averaged time series in each level are hierarchically combined until the final average is achieved. Otherwise, another method – sequential hierarchical averaging – has been suggested, as shown in Figure 1 (b). Many subsequent publications inherit this method under the restriction of having the power of two time series data. In this paper, we will show that the proposed method in [8] does not have associative property as claimed.

(a) (b)

Fig. 1. Two averaging method – (a) balanced hierarchical averaging and (b) sequential hierarchical averaging

The rest of the paper is organized as follows. Section 2 explains some of important background involving shape averaging. Section 3 reveals the problems with current shape averaging method by extensive set of experiments. Finally, in section 4, we conclude with some discussion of potential causes of these inaccuracies, and suggest possible treatment to shape-based time series averaging problem.

2 Background

In this section, we provide brief details of Dynamic Time Warping (DTW) distance measure, its properties, time series averaging using DTW.

2.1 Distance Measurement

Distance measure is extensively used in finding the similarity/dissimilarity between time series. The two well known measures are Euclidean distance metric and DTW distance measure. As a distance metric, it must satisfy the four properties – symmetry, self-identity, non-negativity, and triangular inequality.

A distance measure, however, does not need to satisfy all the properties above. Specifically, the triangular inequality does not hold for the DTW distance measure, which is an important key to the explanation why we have such a hard time in shape averaging using Dynamic Time Warping.

2.2 Dynamic Time Warping Distance

DTW [15] is a well-known similarity measure based on shape. It uses dynamic programming technique to find all possible paths, and selects the one with the minimum distance between two time series. To calculate the distance, it creates a distance matrix, where each element in the matrix is cumulative distance of the minimum of three surrounding neighbors. Suppose we have two time series, a sequence Q of length n ($Q = q_1, q_2, \ldots, q_i, \ldots, q_n$) and a sequence C of length m ($C = c_1, c_2, \ldots, c_j, \ldots, c_m$). First, we create an n-by-m matrix where every (i, j) element of the matrix is the cumulative distance of the distance at (i, j) and the minimum of three neighbor elements, where $0<i \leq n$ and $0<j \leq m$. We can define the (i, j) element as:

$$e_{ij} = d_{ij} + \min\{e_{(i-1)(j-1)}, e_{(i-1)j}, e_{i(j-1)}\} \tag{1}$$

where $d_{ij} = (c_i - q_j)^2$ and e_{ij} is (i, j) element of the matrix which is the summation between the squared distance of q_i and c_j, and the minimum cumulative distance of three elements surrounding the (i, j) element. Then, to find an optimal path, we choose the path that has minimum cumulative distance at (n, m), which is defined as:

$$D_{DTW}(Q,C) = \min_{\forall w \in P}\left\{\sqrt{\sum_{k=1}^{K} d_{w_k}}\right. \tag{2}$$

where P is a set of all possible warping paths, and w_k is (i, j) at k^{th} element of a warping path and K is the length of the warping path.

2.3 Dynamic Time Warping Averaging

Shape averaging exploits DTW distance [8] to find the appropriate mapping for an average. More specifically, the algorithm needs to create a DTW distance matrix and find an optimal warping path. After the path is found, an averaged time series is calculated along this path by using the index (i, j) of each data point w_k on the warping path, which corresponds to the data points q_i and c_j on the time series $Q = q_1, q_2, \ldots, q_i, \ldots, q_n$ and $C = c_1, c_2, \ldots, c_j, \ldots, c_m$, respectively. An optimal warping path W with length K is defined as

$$W = w_1, w_2, \ldots, w_k, \ldots, w_K \tag{3}$$

In addition, w_k, which is mapped with index (i, j), is calculated by the mean value between time series whose indices are i and j. Note that in query refinement, where two time series may have different weights, weight α_Q for a sequence Q and weight α_C for a sequence C, the equation above may then be simply generalized according to the desired weight below

$$w_k = \frac{(\alpha_Q \cdot q_i + \alpha_C \cdot c_j)}{\alpha_Q + \alpha_C} \tag{4}$$

3 Experiment Evaluation

To validate our hypotheses, we set up 4 experiments to disprove the claims of the current shape averaging method. The first experiment tests whether reordering of the sequences will have any effect on the averaged result. The second experiment tests whether DTW averaging of two time series will give the real average. The third experiment tests whether the averaged result is in fact at the center of all original time series. Finally, in the fourth experiment, we test our overall hypotheses by running *k*-means clustering and demonstrate its failure in returning meaningful results.

3.1 Does Reordering Make Any Differences?

This experiment tests whether reordering of the sequences of balanced hierarchical averaging will affect the final averaged time series. According to [8], they claim the associative property under 2^n data constraint, and explicitly state that no matter how we rearrange the data, it will not make any difference in the final averaged outcome.

To show the associative property, we use the Cylinder-Bell-Funnel (CBF) [17], Leaf, Face, Gun, and ECG dataset, from the UCR time series data mining archive [http://www.cs.ucr.edu/~eamonn/time_series_data/]. The well-known 3-class CBF dataset contains 64 instances with the length of 128 data points. The last 3 datasets are multimedia data transformed into time series [16]. The Face dataset contains 112 total normalized instances of 350 data points. The Leaf dataset contains 442 instances of rescaled lengths of 150 data points. Gun dataset has two classes, with 100 instances each, and each instance has the same length of 150 data points. ECG dataset consists of two different heart-pulse classes; each class contains 28 instances with normalized length of 205 data points. Examples of CBF, Leaf, Face, Gun, and ECG data show in Figures 2, 3, 4, 5, and 6.

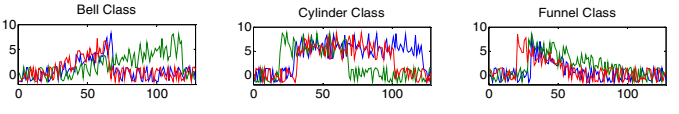

Fig. 2. Examples of CBF data with variations in time axis, i.e. the onset and ending positions

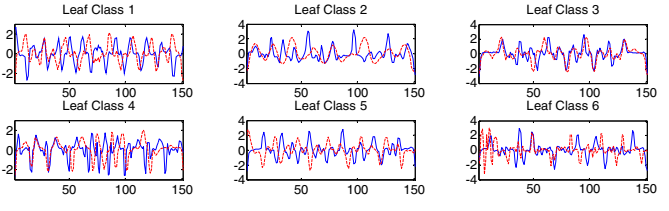

Fig. 3. Examples of six species of Leaf data using time series representation

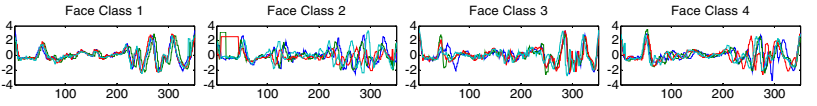

Fig. 4. Examples of six different Face classes

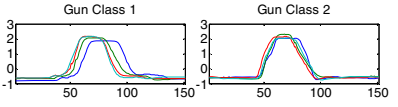

Fig. 5. Examples of Gun data

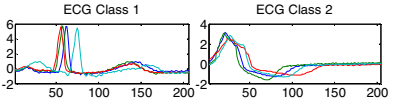

Fig. 6. Examples of ECG data

Even when $n = 1$, we cannot guarantee the commutative property of the averaging method, i.e., $DTW\_Avg(Q,C)$ may or may not give the same result as $DTW\_Avg(C,Q)$, though its symmetric property will give the same DTW distance. When n is larger than one, we want to test whether shuffling the sequences would affect the (balanced hierarchical) averaged result. We first test by averaging only instances within their own class. We compute the distance of *every possible pairings*, then reshuffle the data and repeat the computation (100 runs). We then compare whether the distances among all averaged results from each variation using DTW distance are in fact equal. It is very surprising to see that the averaged time series from each run do not have the same shape, giving the discrepancy among each of the averaging results from different runs much larger than zero. The result are shown in Table 1.

Table 1. Mean and standard deviation of discrepancy distance

	CBF	Leaf	Gun	ECG	Face
Discrepancy distance	227.95±17.23	99.32±10.66	142.50±13.31	6.17±0.90	24.58±2.49

3.2 Correctness of DTW Averaging Between Two Time Series

In this experiment, we demonstrate that the averaged time series, when comparing back to the two original time series, does not have the same distance. Our general intuitive hypothesis is that if we average two time series, the averaged result should *equally* contain characteristics from both original time series. To examine this, we compute the DTW distance from the averaged result back to *both* original time series, and we *should* get the same distance. If this property does not hold, the large number of data mining algorithms that have used this averaging method would probably have worked incorrectly, especially in [8] itself (balanced hierarchical averaging by pairing of 2^n time series). For evaluation, mean percentage errors in all possible pairs in each dataset are computed. Suppose we have two original time series, Q and C, and their resulting averaged time series X. The percentage error is defined as

$$PercentageError(Q,C,X) = \frac{\left| D_{DTW}(Q,X) - D_{DTW}(C,X) \right|}{\max\{D_{DTW}(Q,X), D_{DTW}(C,X)\}} \quad (5)$$

Table 2 shows the experiment results with mean and the standard deviation of percentage errors in each dataset. Note that the percentage error should be 0%.

Table 2. The percentage error from the average results in each dataset

	CBF	Leaf	Gun	ECG	Face
Discrepancy distance	227.95±17.23	99.32±10.66	142.50±13.31	6.17±0.90	24.58±2.49

3.3 Can Cluster Center Drift Out of the Cluster?

In this experiment, we demonstrate an undesirable phenomenon that could happen when we average more than two time series using DTW. We first test on the simplest case where there are only 3 objects to average. We combine hierarchical averaging and sequential averaging methods proposed by [8], to make sure that all three time series are used in the averaging process. We then determine whether the averaged result is in the middle of the group. The example is shown in Figure 7 (a). We first average data, *A*-*B*-*C* from all three data points – *A*, *B*, and *C*. In Figure 7 (b), we calculate average results between each pair of the data points – *A*-*B*, *A*-*C*, and *B*-*C*. If DTW distances between: *A*-*B* and *C* is less than that between *A*-*B*-*C* and *C*, *A*-*C* and *B* is less than that between *A*-*B*-*C* and *B*, or *B*-*C* and *A* is less than that between *A*-*B*-*C* and *A*, then that means the averaged result, *A*-*B*-*C*, is not in the center of the data points. Figure 7 (c) shows the averaged result satisfying the above assumption, but Figure 7 (d) shows the averaged result violating the assumption.

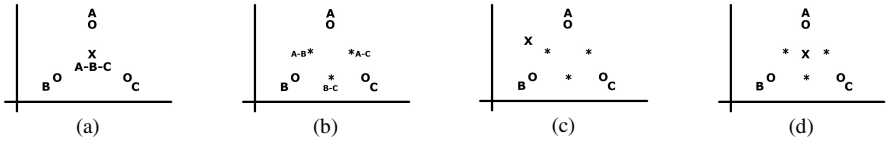

Fig. 7. The cluster center from DTW average, that may drift from the actual cluster center

Table 3 shows the percentage of occurrences that averaged results from all possible three time series that are outside the data group (unsatisfied averaging). Note that the unsatisfied averaging percentage must be 0% to verify correctness of this averaging method.

Table 3. Percentages of average results that are outside the group

	CBF	Leaf	Gun	ECG	Face
Unsatisfied averaged result (%)	0.008%	0.031%	8.936%	23.521%	0.006%

3.4 Failure in *K*-Means Clustering with DTW

This last experiment shows the use of *k*-means clustering using DTW distance to find the cluster center (shape averaging), comparing with an unproblematic *k*-medoids clustering with DTW distance. In this experiment, we will show that if *k*-means

method is used in clustering, there is a high probability of failure (and that is probably why we do not see much of *k*-means clustering with DTW averaging in the literature. We will show this by reporting the average number of iteration up to the point where *k*-means fails, which is when the averaged cluster center happens to drift outside of the cluster, as we discussed in Experiment 3. We run *k*-means clustering with the same datasets. In each dataset, we choose *k* to be its actual numbers of class. For clearer evaluation, we compare the results with the number of iteration obtained using the *k*-medoids methods (which always succeed). We run the experiment 1,000 times for each dataset. Table 4 shows the mean and standard deviation of the number of iteration when *k*-means fails to give meaningful clustering results for each dataset.

Table 4. The mean and standard deviation of number of iteration when *k*-means fails compares to *k*-medoids successes for each dataset

	CBF	**Leaf**	**Gun**	**ECG**	**Face**
Failure: iteration (*k*-means)	2.16±1.13	1.32±0.34	5.16±1.71	1.76±0.76	1.72±0.83
Success: iteration (*k*-medoids)	3.87±0.94	4.19±0.90	4.06±0.93	2.50±0.51	3.61±0.72

4 Discussion, Conclusion, and Future Works

In search of the remedies, we can categorize the problems into three parts, i.e., a distance measure, an averaging method, and dataset properties. First, since DTW is the distance measure that has no triangular inequality property, the averaged time series may not be the actual mean because DTW cannot guarantee the position of averaged result in Euclidean space. Second, in finding a new the averaging method, we suggest that a new averaging method should satisfy various criteria in our proposed experiments. And third, to satisfy triangular inequalities, it also depends on the properties of the data at hands (generally, only a handful of data within a dataset would violate the triangular inequalities). It is possible to first split the data into groups that triangular inequalities hold within. We can simply find the DTW average for each group, and then finally merge those averages together.

In conclusion, we have empirically demonstrated various counterexamples to current shape averaging method using Dynamic Time Warping distance. From these experiments' findings, we have confirmed that the current DTW averaging is inaccurate and should not be used as a subroutine where shape averaging is needed due to lacks of several properties discussed earlier. We conjecture that the reason to this undesirable phenomenon is the triangular inequality that DTW also lacks of. Therefore, DTW averaging cannot guarantee the correctness of the averaging result.

In this paper, we intend to make a first attempt in pointing out some misunderstanding and misuse of current DTW averaging method. As our future work, from these findings, we will investigate how these problems can be resolved and come up with a remedy in accurately averaging shape-based time series data.

References

1. Abdulla, W.H., Chow, D., and Sin, G. Cross-words reference template for DTW-based speech recognition systems. In Proc. of TENCON 2003 (2003)
2. Bagnall, A. and Janacek, G. 2005. Clustering Time Series with Clipped Data. Mach. Learn. 58 (2005) 151-178
3. Boudaoud, S., Rix, H., and Meste, O. Integral shape averaging and structural average estimation. IEEE Trans. on Signal Processing, vol.53, no10 (2005) 3644-3650
4. Bradley, P. S., and Fayyad, U.M. Refining Initial Points for K--Means Clustering. In Proceedings of the 15th Int'l Conference on Machine Learning (1998) 91-99
5. Caiani, E.G., Porta, A., Baselli, G., Turiel, M., Muzzupappa, S., Pieruzzi, F., Crema, C., Malliani, A. and Cerutti, S. Warped-average template technique to track on a cycle-bycycle basis the cardiac filling phases on left ventricular volume. IEEE Computers in Cardiology (1998)
6. Corradini, A. Dynamic Time Warping for Off-Line Recognition of a Small Gesture Vocabulary. In Proc. of the IEEE ICCV Workshop on Ratfg-Rts. Washington, DC (2001)
7. Chu, S., Keogh, E., Hart, and D., Pazzani, M. Iterative deepening dynamic time warping for time series. In Proceedings of SIAM International Conference on Data Mining (2002)
8. Gupta, L., Molfese, D.L., Tammana, R., and Simos, P.G. Nonlinear alignment and averaging for estimating the evoked potential. IEEE Trans. on Biomed. Eng. vol.43, no.4 (1996) 348-356
9. Hu, J. and Ray, B. An Interleaved HMM/DTW Approach to Robust Time Series Clustering. IBM T.J. Watson Research Center (2006)
10. Keogh, E. and Smyth, P. An enhanced representation of time series which allows fast classification, clustering and relevance feedback. KDD (1997) 24-30
11. Lange, D.H., Pratt, H., and Inbar, G.F. (1997). Modeling and estimation of single evoked brain potential components. IEEE Trans. on Biomed. Eng. Vol.44 (1997) 791-799
12. Mor-Avi, V., Gillesberg, I.E., Korcarz, C., Sandelski, J., Lang, R.M. Signal averaging helps reliable noninvasive monitoring of left ventricular dimensions based on acoustic quantification. Computers in Cardiology (1994) 21-24
13. Oates, T., Firoiu, L., and Cohen, P.R. Using Dynamic Time Warping to Bootstrap HMM-Based Clustering of Time Series. In Sequence Learning Paradigms, Algorithms, and Applications. Volume 1828 of Lecture Notes in Computer Science (2001) 35-52
14. Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G. Speaker-independent recognition of isolated words using clustering techniques. In Readings in Speech Recognition, CA (1990) 166-179
15. Ratanamahatana, C.A. and Keogh, E. Everything you know about Dynamic Time Warping is Wrong. In Proc. of KDD Workshop on Mining Temporal and Sequential Data (2004)
16. Ratanamahatana., C.A. and Keogh, E. Multimedia Retrieval Using Time Series Representation and Relevance Feedback. In Proc. of 8th ICADL (2005)
17. Saito, N. Local feature extraction and its application using a library of bases. PhD thesis, Yale University (1994)
18. Salvador, S. and Chan, P. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. In Proc. of KDD Workshop on Mining Temporal and Sequential Data (2004)
19. Wilpon, J. and Rabiner, L. A modified K-means clustering algorithm for use in isolated work recognition. IEEE Trans. on Signal Processing. vol.33 (1985) 587-594