

Non-isometric transforms in time series classification using DTW

Tomasz Górecki ¹ Maciej Łuczak ²

¹Faculty of Mathematics and Computer Science
Adam Mickiewicz University

²Department of Civil and Environmental Engineering
Koszalin University of Technology

International Federation of Classification Societies Conference, 2013

Over recent years the popularity of time series has soared. As a consequence there has been a dramatic increase in the amount of interest in querying and mining such data. In particular, many new distance measures between time series have been introduced. We propose a new distance function based on derivatives and transforms of time series. In contrast to well-known measures from the literature, our approach combines three distances: DTW distance between time series, DTW distance between derivatives of time series, and DTW distance between transforms of time series. The new distance is used in classification with the nearest neighbor rule. In order to provide a comprehensive comparison, we conducted a set of experiments, testing effectiveness on 47 time series data sets from a wide variety of application domains. Our experiments show that this new method provides a significantly more accurate classification on the examined data sets.

The use of derivatives in time series classification is not a novelty. Their use with DTW was proposed by Keogh and Pazzani (2001), who called their method Derivative Dynamic Time Warping (DDTW). Our previous work (Górecki, Łuczak (2013)) contains the results of research on DDTW where the derivative is added, while at the same time parameterization involves the participation of function and derivative. As was shown, such an approach gave very good results. Therefore we decided to conduct further research. We were looking for functions other than the derivative which can be used in a similar manner. The choice was mathematical transforms, which are very popular in the classification of time series.

Ultimately, we decided on three real transforms: sine, cosine and Hilbert. Of course, we do not think that it is enough to compare only the distance between the transforms. It seems natural to add transforms as a further element to improve the accuracy of classification. The parametric approach was used, which allows us to choose the impact of each distance on the final distance measure between the time series, and consequently on the quality of the classification. The new distance functions so constructed are used in the nearest neighbor classification method.

Dynamic time warping (DTW)

DTW is an algorithm for measuring similarity between two sequences which may vary in time or speed. The sequences are warped in a nonlinear fashion to match each other. The basic problem that DTW attempts to solve is how to align two sequences in order to generate the most representative distance measure of their overall difference. The DTW algorithm uses a dynamic programming technique to find an optimal match between two sequences of signals which allows for stretched and compressed sections of the sequence. The first step is to compare each point in one signal with every point in the second signal, generating a matrix. The second step is to work through this matrix, starting at the bottom-left corner, and ending at the top-right: for each cell, the cumulative distance is calculated by picking the neighbouring cell in the matrix to the left or beneath with the lowest cumulative distance, and adding this value to the distance of the focal cell. When this process is complete, the value in the top-right hand cell represents the distance between the two signals according to the most efficient pathway through the matrix.

Dynamic time warping (DTW)

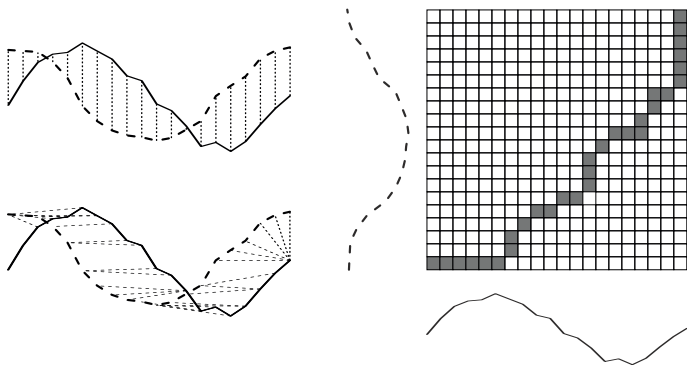


Figure : **Top left**: Two time series which are similar but out of phase produce a large Euclidean distance. **Bottom left**: This can be corrected by DTW's nonlinear alignment. **Right**: To align the signals we construct a warping matrix, and search for the optimal warping path.

Transforms

They are three transforms very popular in technical sciences: the cosine transform, sine transform, and Hilbert transform. All three are non-isometric for DTW distance measure. For a series

$f = \{f(i): i = 1, 2, \dots, n\}$ we have a transform
 $\hat{f} = \{\hat{f}(k): k = 1, 2, \dots, n\}$.

Cosine transform:

$$\hat{f}(k) = \sum_{i=1}^n f(i) \cos \left[\frac{\pi}{n} \left(i - \frac{1}{2} \right) (k - 1) \right].$$

Sine transform:

$$\hat{f}(k) = \sum_{i=1}^n f(i) \sin \left[\frac{\pi}{n} \left(i - \frac{1}{2} \right) k \right].$$

Hilbert transform:

$$\hat{f}(k) = \sum_{\substack{i=1 \\ i \neq k}}^n \frac{f(i)}{k - i}.$$

If dist is a distance measure for two time series f and g , a new distance measure $\widehat{\text{dist}}_{abc}$ is defined by

$$\widehat{\text{dist}}_{abc}(f, g) := a \text{dist}(f, g) + b \text{dist}(f', g') + c \text{dist}(\hat{f}, \hat{g}),$$

where f', g' are discrete (first) derivatives of f, g ; \hat{f}, \hat{g} are transforms; and $a, b, c \in [0, 1]$ are parameters. The discrete derivative of a time series f with length n is defined by

$$f'(i) = f(i) - f(i - 1), \quad i = 2, 3, \dots, n$$

where f' is a time series with length $n - 1$.

Parameters a, b, c are chosen in the tuning phase of a learning process. For an arbitrary distance function dist , we will denote the new distance measure by $\text{DTD}_{\text{dist}}^{\text{trans}}$ (derivative-transform distance), for example $\text{DTD}_{\text{DTW}}^{\text{C}}$ (for the cosine transform).

We don't have to check all values of $a, b, c \in [0, 1]$. We can choose points (a, b, c) on any continuous surface between points $A = (1, 0, 0)$, $B = (0, 1, 0)$, and $C = (0, 0, 1)$. For example, it can be a surface of the triangle with vertices in those points or one eighth of sphere. For simplicity, we choose the triangle. This 3d triangle we map to the 2d triangle with vertices in points $A' = (0, 0)$, $B' = (1, 0)$, and $C' = (\frac{1}{2}, \frac{\sqrt{3}}{2})$. Both triangles we can define in parametrical way:

$$(a, b, c) = A + \alpha \overrightarrow{AB} + \beta \overrightarrow{AC},$$
$$(a', b') = A' + \alpha \overrightarrow{A'B'} + \beta \overrightarrow{A'C'},$$

where (a, b, c) are points of the 3d triangle, (a', b') are points of the 2d triangle, and $\alpha, \beta \in [0, 1]$ are parameters.

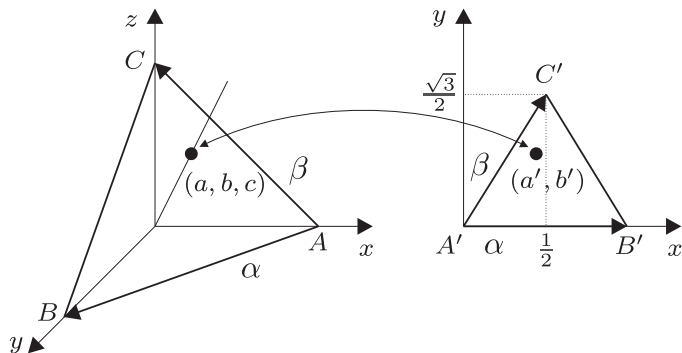


Figure : Dependence of 3D parameters a, b, c and 2D parameters α, β .

Lower bound and triangular inequality

For many distance measures it can be found a lower bound of them. Then the lower bound can be used in the nearest neighbor method to speed up computations. We can also find a lower bound of our new distance measure. If low is a lower bound of a distance dist , then

$$\widehat{\text{low}}_{abc}(f, g) = a \text{low}(f, g) + b \text{low}(f', g') + c \text{low}(\hat{f}, \hat{g})$$

is a lower bound of the distance $\widehat{\text{dist}}_{abc}$.

If the base distance dist is a metric, then the new distance $\widehat{\text{dist}}$ is also a metric. If dist is not a metric but obeys the triangular inequality, then the distance $\widehat{\text{dist}}$ obeys the triangular inequality as well:

$$\widehat{\text{dist}}_{abc}(f, g) \leq \widehat{\text{dist}}_{abc}(f, h) + \widehat{\text{dist}}_{abc}(h, g).$$

We performed experiments on 47 data sets. The data sets originate from the UCR Time Series Classification/Clustering Homepage (Keogh et al. (2011)). To use with the new distance function, we chose one distance measure: DTW and three transforms: sine, cosine, and Hilbert. Thus we have three similarity measures denoted by DTD_{DTW}^S , DTD_{DTW}^C , and DTD_{DTW}^H . For each data set we calculated the classification error rate on a test subset (to learn the model we used a training subset, leave-one-out, 1NN method). We found all parameters using the training subset. We use the cross-validation (leave-one-out) method to find the best pair of parameters α , β in our classifier. If the minimal error rate is the same for more than one pair of parameters α , β we choose the smallest pair (minimizing α first, then β). Finite subsets of parameters α and β are chosen, from 0 to 1 with fixed step 0.01.

$\frac{DD_{DTW} - DTW}{DTW}$	$\frac{DTD_{DTW}^C - DTW}{DTW}$	$\frac{DTD_{DTW}^S - DTW}{DTW}$	$\frac{DTD_{DTW}^H - DTW}{DTW}$
-19.03	-22.72	-22.21	-19.75

Comparing the new distances with the standard DTW we can see a significant reduction in error rate for most data sets. This is especially clearly seen for the mean of relative errors. The average relative error reduction for all data sets is equal to 22.72% for DTD_{DTW}^C , 22.21% for DTD_{DTW}^S , and 19.75% for TD_{DTW}^H . The reduction for DD_{DTW} method is 19.03%, therefore the new distances are slightly better than DD_{DTW}

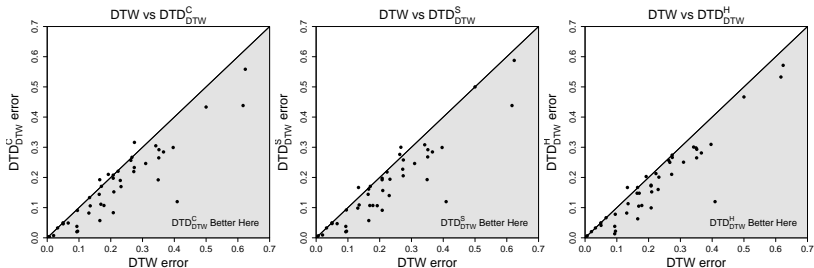


Figure : Comparison of test errors (DTW vs DTD_{DTW}^*).

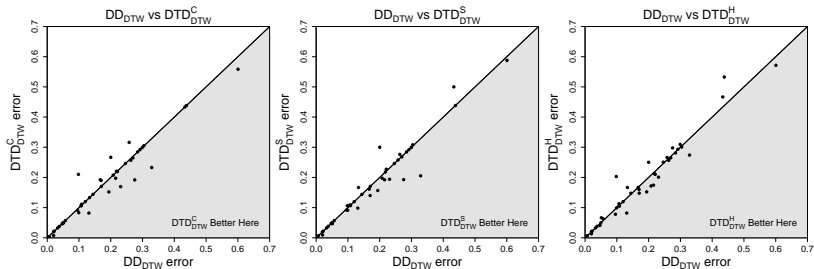


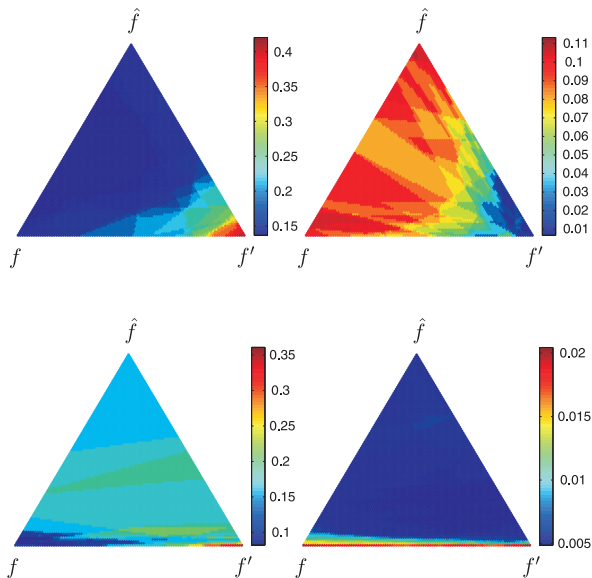
Figure : Comparison of test errors (DD_{DTW} vs DTD_{DTW}^*).



To find differences between the methods we used the Iman and Davenport (1980) test, which is a nonparametric equivalent of ANOVA. Due to the fact that the p -value is equal to 0, we can proceed with the post-hoc test in order to detect significant pairwise differences among all the classifiers. As a post-hoc test we used Bergmann and Hommel (1988) dynamic procedure.

Procedure	Ranks mean	
DTD_{DTW}^S	2.57	*
DTD_{DTW}^C	2.59	*
DTD_{DTW}^H	2.74	*
DD_{DTW}	3.02	*
DTW	4.06	*

Finally, we have two homogeneous disjoint groups of classifiers. The best classifiers are in the first group.

Of course, the interesting question is that of what derivative and transform contributions in the final distance measure are optimal. Could we obtain some arbitrary quantity that determines for all cases that such and such participation will give us the best result of classification? The answer to this question is negative. The optimal share may be zero, average, or that it exclusively should be used. Following Figures present the contribution of each component of the new distance, i.e. the value of f , its derivative f' and its transform \hat{f} , for the example data sets. We can see that the minimal error can be obtained at different points of the triangle, i.e. for different values of parameters a, b, c (α, β).



-  Bergmann, G., Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses, in P. Bauer, G. Hommel and E. Sonnemann (Eds), Multiple Hypotheses Testing, Springer, pp. 110–115.
-  Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E. (2008). Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. In Proc. 34th Int. Conf. on Very Large Data Bases, 1542–1552.
-  Górecki, T., Łuczak, M. (2013). Using derivatives in time series classification. Data Mining and Knowledge Discovery 26(2): 310-331.
-  Keogh, E., Pazzani, M. (2001). Dynamic Time Warping with Higher Order Features. In First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA.
-  Keogh, E., Zhu, Q., Hu, B., Hao, Y., Xi, X., Wei, L. & Ratanamahatana, C. A. (2011). The UCR Time Series Classification/Clustering Homepage: http://www.cs.ucr.edu/~eamonn/time_series_data/