A Hybrid Metric based on Persistent Homology and its Application to Signal Classification

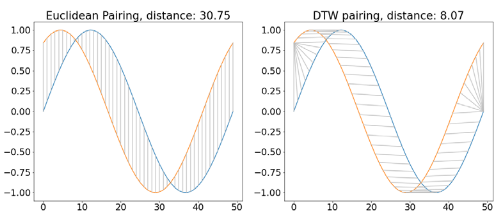
Austin Lawson
Program of Informatics and Analytics, UNC Greensboro

Yu-Min Chung & William Cruse
Department of Mathematics & Statistics, UNC Greensboro

Background

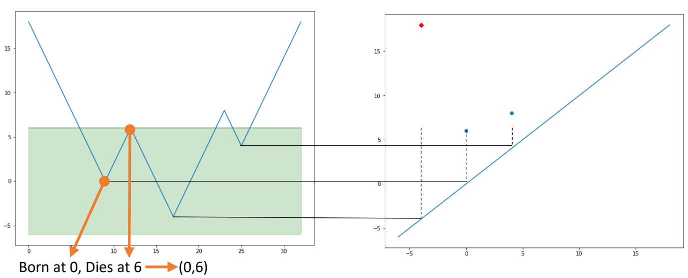
Time Series

Let $[n] = \{0,1,\dots,n-1\}$. A time series is a function $s: [n] \to \mathbb{R}$. We often use s_i in place of $s(i)$. We use two popular distances with time series, the Euclidean ℓ^1-distance, $||s-t||_1 = \sum_{i=0}^{n-1} |s_i - t_i|$ and the Dynamic Time Warping Distance, which we briefly define. A **warping path** w between s and t is a function $w: [N] \to [n] \times [m]$ so that $w(1) = (0,0), w(N) = (n,m)$ and $w_{\{i+1\}} - w_i \in \{(1,0),(1,1),(0,1)\}$ where N is the length of the path from $(0,0)$ to (n,m). We denote the collection of warping path between s and t by $\Omega_{s,t}$. The **cost** of a warping path w is defined by $K(w) = \sum_{i=0}^{N-1} |s_{w(i)} - t_{w(i)}|$. Then the **Dynamic Time Warping Distance** between s and t is defined by $||s-t||_{DTW} = \min_{w \in \Omega_{s,t}} K(w)$.



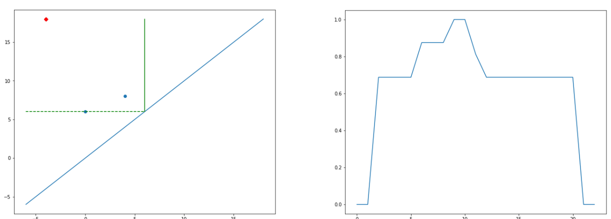
Persistent Homology

Produces a persistence diagram, which is a summary of topological information.



Persistence Curves

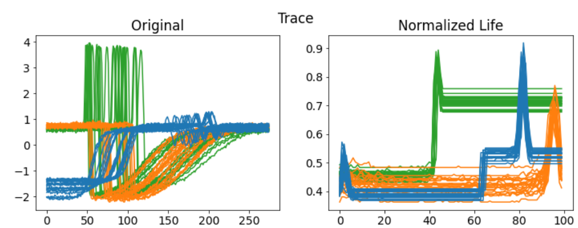
Further summarize persistence diagrams by transformation to a 1-D real function. In short, for any real number x, we consider the points of the diagram lying inside the quadrant whose corner lies at (x,x). We apply some function and a statistic to those points. Specifically, for this work we subtract the birth (x-value) from the death (y-value) and divide by the total of this function over the diagram. Then in each quadrant, we simply sum the function values. This yields the curve on the right below, called the **normalized lifespan curve**, denoted $\mathbf{sl}(s)$ for a time series s.

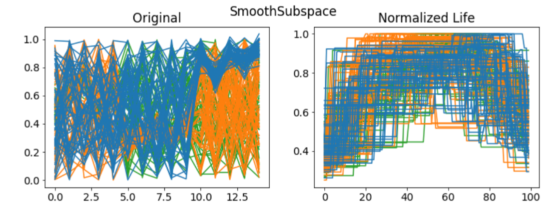


The Hybrid Metric

Persistence curves transforms time series into a time series containing topological information. With our choice in the way we compute persistence, we can easily differentiate time series that behave different "vertically". However, we suffer on time series that differ "horizontally". We can define a metric that accounts for the original space as well as the topological information. For $\alpha \in (0,1]$ and $X \in \{1, DTW\}$ define

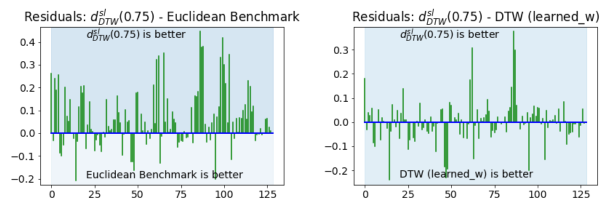
$$d_X^{sl}(\alpha; s,t) = \alpha||s-t||_X + (1-\alpha)||\mathbf{sl}(s) - \mathbf{sl}(t)||_X.$$





Data, Model, and Results

The UCR Time Series Classification Archive contains 128 datasets of time series gathered by different means for different scenarios and contributed by several different research groups. Each dataset varies in train/test size and time series length. The task for each of these datasets is classification, and each dataset has a different number of classes. As recommended by the archive, we tested our metric by using a 1-NN classifier for each dataset and compared with the benchmarks used in UCR. Overall, our metrics dominate the basic Euclidean norm and the hybrid DTW performs slightly better than the base DTW as indicated in the plots below.



Discussion

The methods in this work serve as a pilot to a larger plot. The overarching idea is to use distance to measure difference in topological information. The key here is that our chosen method of persistent homology is not the only option for time series. In future work, we seek to employ delay embeddings and other transformations to extract better topological information. Moreover, the persistence curve framework is rich and contains many more summaries than the one used here.