Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

Time Series Classification in Dissimilarity Spaces

Brijnesh J. Jain and Stephan Spiegel
Berlin Institute of Technology, Germany
jain@dai-labor.de, spiegel@dai-labor.de

Abstract. Time series classification in the dissimilarity space combines the advantages of the dynamic time warping and the rich mathematical structure of Euclidean spaces. We applied dimension reduction using PCA followed by support vector learning on dissimilarity representations to 43 UCR datasets. Results indicate that time series classification in dissimilarity space has potential to complement the state-of-the-art.

1 Introduction

Time series classification finds many applications in diverse domains such as speech recognition, medical signal analysis, and recognition of gestures [2–4]. Surprisingly, the simple nearest-neighbor method together with the dynamic time warping (DTW) distance still belongs to the state-of-the-art and is reported to be *exceptionally difficult to beat* [1, 5, 10]. This finding is in stark contrast to classification in Euclidean spaces, where nearest neighbor methods often merely serve as baseline. One reason for this situation is that nearest neighbor methods in Euclidean spaces compete against a plethora of powerful statistical learning methods. The majority of these statistical learning methods are based on the concept of derivative not available for warping-invariant functions on time series.

The dissimilarity space approach proposed by [7] offers to combine the advantages of the DTW distance with the rich mathematical structure of Euclidean spaces. The basic idea is to first select a set of k reference time series, called prototypes. Then the dissimilarity representation of a time series consists of k features, each of which represents its DTW distance from one of the k prototypes. Since dissimilarity representations are vectors from $\mathbb{R}^k$, we can resort to the whole arsenal of mathematical tools for statistical data analysis. The dissimilarity space approach has been systematically applied to the graph domain using graph matching [6, 9]. A similar systematic study of the dissimilarity space approach for time series endowed with the DTW distance is still missing.

This paper is a first step towards exploring the dissimilarity space approach for time series under DTW. We hypothesize that combining the advantages of both, the DTW distance and statistical pattern recognition methods, can result in powerful classifiers that may complement the state-of-the-art. The proposed approach applies principal component analysis (PCA) for dimension reduction of the dissimilarity representations followed by training a support vector machine (SVM). Experimental results provide support for our hypothesis.

2 Dissimilarity Representations of Time Series

2.1 Dynamic Time Warping Distance

A time series of length n is an ordered sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ with features $x_i \in \mathbb{R}$ sampled at discrete points of time $i \in [n] = \{1, \ldots, n\}$. To define the DTW distance between time series $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n and m, resp., we construct a grid $\mathcal{G} = [n] \times [m]$. A warping path in grid $\mathcal{G}$ is a sequence $\phi = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p)$ consisting of points $\boldsymbol{t}_k = (i_k, j_k) \in \mathcal{G}$ such that

1. $\boldsymbol{t}_1 = (1,1)$ and $\boldsymbol{t}_p = (n,m)$ (boundary conditions)
2. $\boldsymbol{t}_{k+1} - \boldsymbol{t}_k \in \{(1,0), (0,1), (1,1)\}$ (warping conditions)

for all $1 \leq k < p$. The cost of warping $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_m)$ along ϕ is defined by

$$d_\phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{(i,j) \in \phi} (x_i - y_j)^2,$$

where $(x_i - y_j)^2$ is the local transformation cost of assigning features x_i to y_j. Then the distance function

$$d(\boldsymbol{x}, \boldsymbol{y}) = \min_{\phi} d_\phi(\boldsymbol{x}, \boldsymbol{y}),$$

is the dynamic time warping (DTW) distance between $\boldsymbol{x}$ and $\boldsymbol{y}$, where the minimum is taken over all warping paths in $\mathcal{G}$.

2.2 Dissimilarity Representations

Let $(\mathcal{T}, d)$ be a time series space $\mathcal{T}$ endowed with the DTW distance d. Suppose that we are given a subset

$$\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\} \subseteq \mathcal{T}$$

of k reference time series $\boldsymbol{p}_i \in \mathcal{T}$, called prototypes henceforth. The set $\mathcal{P}$ of prototypes gives rise to a function of the form

$$\phi : \mathcal{T} \to \mathbb{R}^k, \quad \boldsymbol{x} \mapsto (d(\boldsymbol{x}, \boldsymbol{p}_1), \ldots, d(\boldsymbol{x}, \boldsymbol{p}_k)),$$

where $\mathbb{R}^k$ is the *dissimilarity space* of $(\mathcal{T}, d)$ with respect to $\mathcal{P}$. The k-dimensional vector $\phi(\boldsymbol{x})$ is the *dissimilarity representation* of $\boldsymbol{x}$. The i-th feature of $\phi(\boldsymbol{x})$ represents the dissimilarity $d(\boldsymbol{x}, \boldsymbol{p}_i)$ between $\boldsymbol{x}$ and the i-th prototype $\boldsymbol{p}_i$.

2.3 Learning Classifiers in Dissimilarity Space

Suppose that

$$\mathcal{X} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\} \subseteq \mathcal{T} \times \mathcal{Y}.$$

is a training set consisting of n time series $\boldsymbol{x}_i$ with corresponding class labels $y_i \in \mathcal{Y}$. Learning in dissimilarity space proceeds in three steps: (1) select a

suitable set of prototypes $\mathcal{P}$ on the basis of the training set $\mathcal{D}$, (2) embed time series into the dissimilarity space by means of their dissimilarity representations, and (3) learn a classifier in the dissimilarity space according to the empirical risk minimization principle.

The performance of a classifier learned in dissimilarity spaces crucially depends on a proper dissimilarity representation of the time series. We distinguish between two common approaches:

1. *Prototype selection*: construct a set of prototypes $\mathcal{P}$ from the training set $\mathcal{X}$.
2. *Dimension reduction*: perform dimension reduction in the dissimilarity space.

There are numerous strategies for prototype selection. Naive examples include all elements of the training set $\mathcal{X}$ and sampling a random subset of $\mathcal{X}$. For more sophisticated selection methods, we refer to [8]. Dimension reduction of the dissimilarity representation includes methods such as, for example, principal component analysis (PCA).

3 Experiments

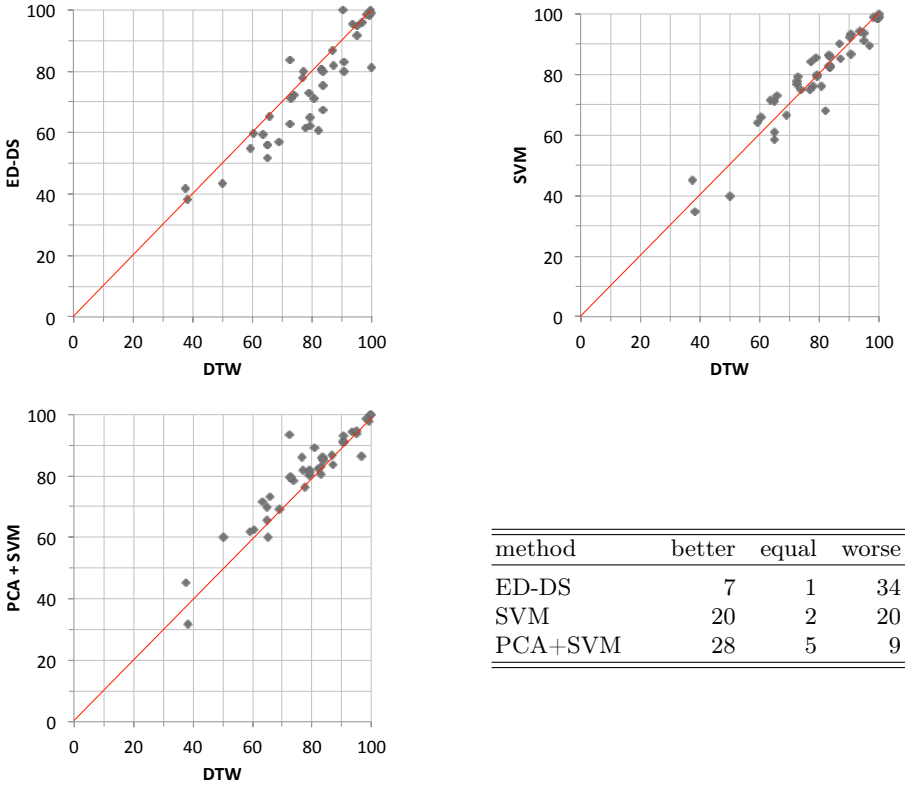
The goal of this experiment is to assess the performance of the following classifiers in dissimilarity space: (1) nearest neighbor using the Euclidean distance (ED-DS), (2) support vector machine (SVM), and (3) principal component analysis on dissimilarity representations followed by support vector machine (PCA+SVM).

3.1 Experimental Protocol

We considered 43 datasets from the UCR time series datasets [4], each of which comes with a pre-defined training and test set. For each dataset we used the whole training set as prototype set. To embed the training and test examples into a dissimilarity space, we computed their DTW distances to the prototypes.

We trained all SVMs with RBF-kernel using the embedded training examples. We selected the parameters γ and C of the RBF-kernel over a two-dimensional grid with points $(\gamma_i, C_j) = (2^i, 2^j)$, where i, j are 30 equidistant values from $[-10, 10]$. For each parameter configuration (γ_i, C_j) we performed 10-fold cross-validation and selected the parameters (γ_*, C_*) with the lowest average classification error. Then we trained the SVM on the whole embedded training set using the selected parameters (γ_*, C_*). Finally, we applied the learned model to the embedded test examples for estimating the generalization performance.

For PCA+SVM we performed dimension reduction using PCA prior training of the SVM. We considered the q first dimensions with highest variance, where $q \in \{1, 1+a, 1+2a, \ldots, 1+19a\}$ with a being the closest integer of $k/20$ and k is the dimension of the dissimilarity space. For each q, we performed hyper-parameter selection for the SVM as described above. We selected the parameter configuration (q_*, γ_*, C_*) that gave the lowest classification error. Then we applied PCA on the whole embedded training set, retained the first q_* dimensions and trained the SVM on the embedded training set after dimension reduction.



method	better	equal	worse
ED-DS	7	1	34
SVM	20	2	20
PCA+SVM	28	5	9

Fig. 1. Scatter plots of accuracy of DTW against dissimilarity space methods.

Finally we reduced the dimension of the embedded test examples and applied the learned model.

3.2 Results

Figure 1 shows the scatter plots of predictive accuracy of the nearest neighbor using DTW against all three dissimilarity space methods and Table 1 shows the error rates of all classifiers for each dataset.

The first observation to be made is that the dissimilarity space endowed with the Euclidean space is less discriminative than the time series space endowed with the DTW distance. As shown by Figure 1, nearest neighbor (NN) with DTW performed better than the ED-DS classifier in 34 out of 42 cases. Since the DTW distance is non-Euclidean, dissimilarity spaces form a distorted representation of the time series space in such a way that neighborhood relations are not preserved. In most cases, these distortions impact classification results negatively, often by a large margin. In the few cases where the distortions improve classification

Data	DTW	ED-DS	SVM	PCA+SVM
50words	31.0	42.9	33.4	31.0 (+0.0)
Adiac	39.6	40.2	34.0	37.6 (+5.1)
Beef	50.0	56.7	60.0	40.0 (+20.0)
CBF	0.3	0.2	1.4	0.2 (+32.7)
ChlorineConcentration	35.2	48.3	28.9	30.2 (+14.3)
CinC ECG Torso	34.9	44.0	41.4	39.8 (-14.0)
Coffee	17.9	39.3	32.1	17.9 (+0.0)
Cricket X	22.3	38.5	24.1	23.6 (-5.8)
Cricket Y	20.8	37.9	20.0	19.7 (+5.1)
Cricket Z	20.8	34.9	20.8	18.2 (+12.5)
DiatomSizeReduction	3.3	4.2	10.8	13.7 (-315.9)
ECG	23.0	20.0	16.0	18.0 (+21.7)
ECGFiveDays	23.2	22.4	24.9	14.1 (+39.4)
Face (all)	19.2	28.9	23.8	10.9 (+43.0)
Face (four)	17.1	19.3	13.6	17.0 (+0.0)
FacesUCR	9.5	17.1	13.4	9.0 (+5.6)
Fish	16.7	32.6	17.7	14.3 (+14.5)
Gun-Point	9.3	20.0	6.7	8.7 (+7.1)
Haptics	62.3	58.4	54.9	54.9 (+11.9)
InlineSkate	61.6	61.8	65.5	68.4 (-11.0)
ItalyPowerDemand	5.0	8.4	6.5	6.3 (-26.3)
Lighting 2	13.1	18.0	14.8	16.4 (-25.1)
Lighting 7	27.4	37.0	23.3	20.5 (+25.0)
Mallat	6.6	4.6	5.6	5.5 (+17.3)
Medical Images	26.3	27.9	24.9	21.7 (+17.4)
MoteStrain	16.5	24.8	17.2	14.1 (+14.8)
Olive Oil	13.3	13.3	10.0	13.3 (+0.0)
OSU Leaf	40.9	45.0	36.0	38.0 (+7.1)
SonyAIBORobotSurface	27.5	16.3	22.3	6.7 (+75.8)
SonyAIBORobotSurface II	16.9	19.4	17.4	19.3 (-14.2)
Swedish Leaf	21.0	27.2	14.4	18.7 (+10.9)
Symbols	5.0	5.3	8.7	5.3 (-6.5)
Synthetic Control	0.7	1.7	1.3	2.0 (-185.7)
Trace	0.0	1.0	1.0	0.0 (+0.0)
TwoLeadECG	9.6	18.7	7.7	7.1 (+25.9)
TwoPatterns	0.0	18.7	0.0	0.0 (+0.0
uWaveGestureLibrary X	27.3	28.9	20.8	20.6 (+24.6)
uWaveGestureLibrary Y	36.6	40.5	28.6	28.5 (+22.2)
uWaveGestureLibrary Z	34.2	34.6	27.0	26.9 (+21.4)
Wafer	2.0	1.5	1.1	1.5 (+26.2)
WordsSynonyms	35.1	43.9	39.0	34.3 (+2.2)
yoga	16.4	20.0	14.0	14.8 (+9.6)

Table 1. Error rates in percentages. Numbers in parentheses show percentage improvement of PCA+SVM with respect to DTW.

results, the improvements are only small and could also be occurred by chance due to the random sampling of the training and test set.

The second observation to be made is that the SVM using all prototypes complements NN+DTW. Better and worse predictive performance of both classifiers is balanced. This shows that powerful learning algorithms can partially compensate for poor representations.

The third observation to be made is that SVM+PCA outperformed all other classifiers. Furthermore, SVM+PCA is better than NN+DTW in 28 and worse in 9 out of 42 cases. By reducing the dimension using PCA, we obtain better dissimilarity representations for classification. Table 1 highlights relative improvements and declines of PCA+SVM compared to NN+DTW with $\pm 10\%$ or more in blue and red color, respectively. We observe a relative change of at least $\pm 10\%$ in 27 out of 43 cases. This finding supports our hypothesis that learning on dissimilarity representations complements NN+DTW.

4 Conclusion

This paper is a first step to explore dissimilarity space learning for time series classification under DTW. Results combining PCA with SVM on dissimilarity representations are promising and complement nearest neighbor methods using DTW in time series spaces. Future work aims at exploring further elastic distances, prototype selection, dimension reduction, and learning methods.

References

1. G.E. Batista, X. Wang, and E.J. Keogh. A Complexity-Invariant Distance Measure for Time Series. *SIAM International Conference on Data Mining*, 11:699–710, 2011.
2. T. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.
3. P. Geurts. Pattern extraction for time series classification. *Principles of Data Mining and Knowledge Discovery*, pp. 115–127, 2001.
4. E. Keogh, Q. Zhu, B. Hu, Y. Hao., X. Xi, L. Wei, and C. A. Ratanamahatana. The UCR Time Series Classification/Clustering Homepage: `www.cs.ucr.edu/~eamonn/time_series_data/`, 2011.
5. J. Lines and A. Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 2014.
6. L. Livi, A. Rizzi, and A. Sadeghian. Optimized dissimilarity space embedding for labeled graphs. *Information Sciences*, 266:47–64, 2014.
7. E. Pekalska and R.P.W. Duin *The Dissimilarity Representation for Pattern Recognition.* World Scientific Publishing Co., Inc., 2005.
8. E. Pekalska, R.P.W. Duin, and P. Paclik. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2): 189–208, 2006.
9. K. Riesen and H. Bunke. Graph classification based on vector space embedding. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(6):1053–1081, 2009.
10. X. Xi, E. Keogh, C. Shelton, L. Wei, and C.A. Ratanamahatana. Fast time series classification using numerosity reduction. *International Conference on Machine Learning*, pp. 1033–1040, 2006.