Transfer Learning for Time Series Classification in Dissimilarity Spaces

Stephan Spiegel

DAI Lab, Berlin Institute of Technology
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
spiegel@dai-lab.de
http://www.dai-lab.de/en

Abstract. For many time series classification problems the amount of labeled data is insufficient to achieve satisfying accuracy. However, there exists an abundance of unlabeled time series data, which does not follow the same class labels or generative distribution as the labeled data. We propose to employ this unlabeled data to improve the performance of the supervised learning task, which is also known as transfer learning. In this work use arbitrary UCR time series or random sequences to embed a given set of labeled data into dissimilarity space, leading to enriched feature representations that facilitate statistical learning procedures. Our results show that transfer learning increases the accuracy of time series classification in dissimilarity spaces, which in turn has been shown to outperform the popular 1NN-DTW time series classification approach.

Keywords: Transfer Learning, Dissimilarity Spaces, Time Series

1 Introduction

In time series mining, the Dynamic Time Warping (DTW) distance is a commonly and widely used dissimilarity measure [14]. Its popularity and widespread use are owing to the fact that, in contrast to Euclidean distance (ED), the DTW distance works well for time series with local scaling invariance [1]. The popular combination of the 1-Nearest-Neighbor (1NN) classifier with the DTW distance has been shown to achieve high classification accuracy on time series from various application domains [4].

The 1NN-DTW approach is very intuitive since humans typically compare objects by means of analogies in their structure [6]. But, structural descriptions (such as local scalings captured by DTW) do not match well with statistical learning procedures, which are most powerful for vectorial object representations [5]. Dissimilarity spaces are a promising way to combine structural and statistical pattern recognition approaches [12], where structural descriptions are used to compare objects, leading to a set of pairwise dissimilarities from which vectors can be derived for the purpose of statistical learning [5]. Early work on graphs [2, 15] has shown that the dissimilarity space approach has significant potential to outperform classifiers that directly operate in the graph domain.

Recent work on time series classification [7–9] has proposed to consider DTW distances as feature vectors for standard machine learning methods. The results [7–9] have shown that time series classification in dissimilarity spaces (using the SVM model) is superior to the 'exceptionally hard to beat' 1NN-DTW approach [1, 10, 17]. To furthermore increase the classification accuracy of the dissimilarity space approach, it has been proposed to add cDTW and SAX distances to the feature vectors [9] as well as to extract mutually independent features by means of PCA [7, 8]. Related work [3, 16] has confirmed that the SVM model achieves better generalization performance with prior feature extraction.

In this work, we aim at extending the idea of 'time series classification in dissimilarity spaces' [7, 8] by means of transfer learning [11]. More precisely, we employ self-taught learning, where unlabeled data is used for the supervised classification task [13]. Self-taught learning does not assume that the unlabeled data follows the same class labels or generative distribution, making it widely applicable to many practical learning problems. In this study, we employ a large number of unlabeled temporal sequences and randomly generated processes to embed a given set of time series and to improve their classification.

The rest of the paper is structured as follows. Section 2 introduces our proposed approach. Section 3 presents empirical results, which are discussed in Section 4. Finally, we conclude with future work in Section 5.

2 Approach

In supervised learning, we are usually given a labeled dataset $\mathbb{D} = \{\mathbb{X}, \mathbb{Y}\}$ that consists of x training examples $\mathbb{X} = X_1, \ldots, X_x$ and y test examples $\mathbb{Y} = Y_1, \ldots, Y_y$. The goal is to predict the correct labels for all test examples $\mathbb{Y}$ by generalizing from our training examples $\mathbb{X}$. Most time series classification tasks [4] assume that the test and training examples $X_1, \ldots, X_x, Y_1, \ldots, Y_y \in \mathbb{R}^n$ have the same length or number of dimensions n, although this is no necessary requirement for pairwise (dis)similarity comparisons.

In earlier work [7, 8] we have proposed to solve time series classification in dissimilarity spaces. For this purpose, we have used a selected subset of α training examples $\mathbb{A} \subseteq \mathbb{X}$ to embed our original dataset $\mathbb{D}$ into a new feature space. Given a time series $T \in \mathbb{D}$ we obtain its feature representation $\big(d(T, A_1), \ldots, d(T, A_\alpha)\big)$ by computing the dynamic time warping distance $d(\cdot, \cdot)$ to all training examples contained in $\mathbb{A} = \{A_1, \ldots, A_\alpha\}$. The crux of dissimilarity spaces is that the derived feature vectors reside in Euclidean space and can be used by powerful statistical learning procedures.

In this work aims at studying transfer learning for time series classification in dissimilarity spaces. More precisely, we evaluate the classification accuracy of dissimilarity spaces that were constructed from a set of β unlabeled time series $\mathbb{B} = \{B_1, \ldots, B_\beta\}$ that do not follow the same class labels and generative distribution as the labeled data $\mathbb{D}$. Given a time series $T \in \mathbb{D}$ we now obtain its feature representation $\big(d(T, B_1), \ldots, d(T, B_\beta)\big)$ by computing the dynamic time warping distance $d(\cdot, \cdot)$ to all examples in $\mathbb{B}$.

Of course, we can also concatenate the feature vectors of dissimilarity spaces that were constructed from different datasets. For example, we can use a subset of labeled training examples $\mathbb{A}$ as well as some unlabeled time series $\mathbb{B}$ plus a set of random sequences $\mathbb{H}$ to embed a time series $T \in \mathbb{D}$ into the following dissimilarity space: $\big(d(T, A_1), \ldots, d(T, A_\alpha), d(T, B_1), \ldots, d(T, B_\beta), d(T, H_1), \ldots, d(T, H_\eta)\big)$. In that way, we combine available knowledge from various domains, which can be transferred to our supervised learning task. Please note that we never use the test set for embedding in order to keep it independent from the training set.

Table 1. Dissimilarity Matrices

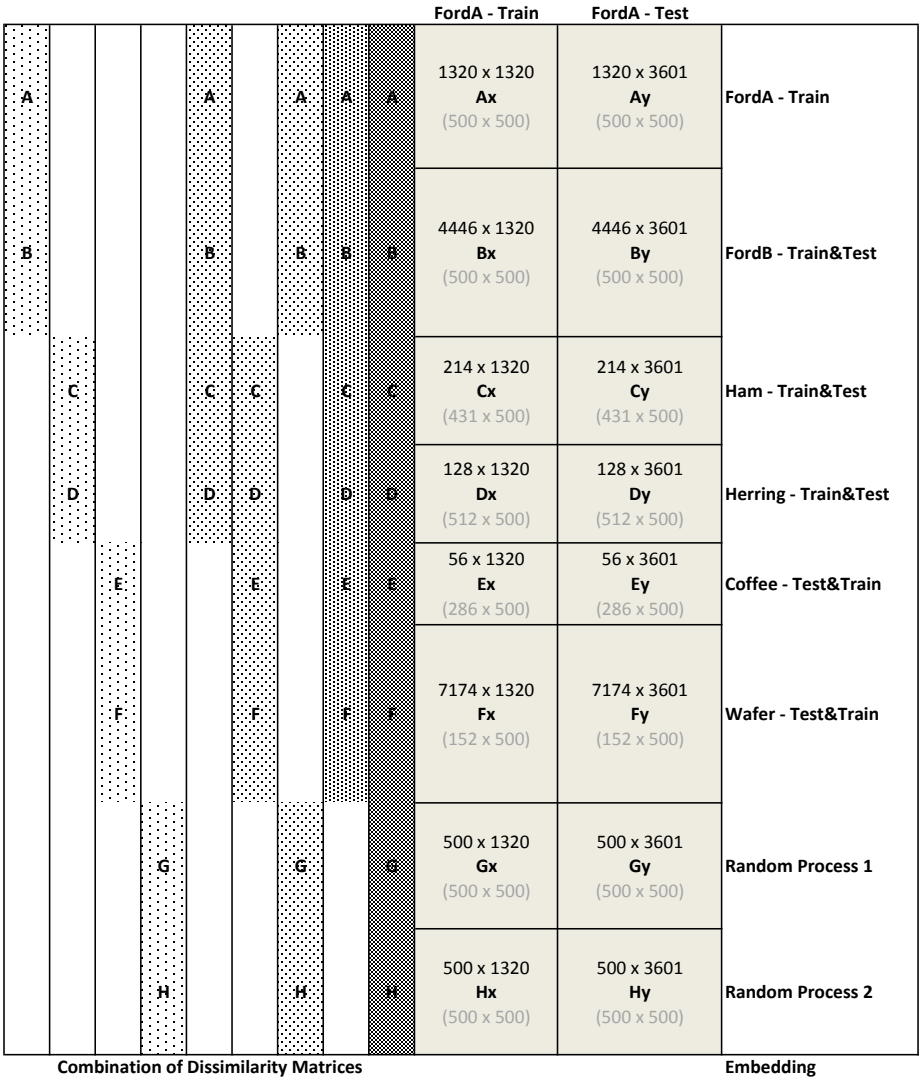

Table 1 illustrates our approach for the *FordA* dataset, which consists of 1320 training and 3601 test time series of length 500. If we embed the *FordA* dataset with all 1320 training examples then we obtain two dissimilarity matrices Ax and Ay of size 1320×1320 and 1320×3601. The columns in Ax and Ay represent our generated feature vectors for the $FordA$ training and test examples respectively. Note that each entry in Ax and Ay corresponds to a time series distance, which is the result of finding an optimal path in a 500×500 warping matrix.

Table 1 furthermore illustrates the dissimilarity matrices for an embedding with two sets of random processes as well as with the *FordB*, *Ham*, *Herring*, *Coffee*, and *Wafer* time series [4], using both training and test examples. For each embedding, we denote the size of the resulting dissimilarity matrices (in black color) as well as the size of the corresponding warping matrices (in brackets and gray color). Given all the illustrated dissimilarity matrices (A to H) we can either consider their feature vectors individually or combine their column vectors to a new feature representations. The left side of Table 1 shows the combinations that we evaluate in Section 3.

3 Results

The goal of our evaluation is to assess the time series classification accuracy in consideration of various different dissimilarity spaces, which were constructed from the combinations of labeled training examples, unlabeled time series, and random sequences.

In our experiments we consider the complementary datasets *Toe1* and *Toe2* as well as *FordA* and *FordB*, which are part of the UCR time series archive [4]. For each of the four datasets we assess the classification accuracy for an embedding with the corresponding training set, the respective complementary dataset, arbitrary unlabeled time series, and random (auto-regressive) processes. Furthermore, we assess the classification accuracy for different combinations of the resulting feature vectors, as illustrated by our example in Table 1. The individual or combined feature vectors can subsequently be used as an input for standard statistical learning procedures. In our experiments we employ a linear SVM (using the quadratic programming algorithm of the Matlab optimization toolbox) to solve the classification problem.

Table 2 shows the classification errors for the traditional 1-Nearest-Neighbor classifier (with ED and DTW) in comparison to our proposed approach, using transfer learning in dissimilarity spaces. More precisely, we present classification results for various different dissimilarity spaces that were constructed by means of individual datasets or their combination. In Table 2, an embedding with a combination of different time series is symbolized by a sequence of capital letters, where each letter represent an individual dataset. For instance, the sequence AB describes the combination of the label training examples A and the corresponding complementary dataset B. Please note that the classification performance does not depend on the ordering of the combined datasets and does not change for different permutations of the constructed feature vectors.

Furthermore, Table 2 presents the classification results for an embedding with arbitrary datasets (*C=Ham*, *D=Herring*, *E=Coffee*, and *F=Wafer* from the UCR time series archive [4]) as well as the two sets of auto-regressive processes (*G=Random Process 1* and *H=Random Process 2* that were generated by two different parameters settings)[1].

The results in Table 2 shows how transfer learning from complimentary, arbitrary, or random time series influences the classification error. We discuss our interpretation of the empirical results in Section 4.

Table 2. Classification Errors

1NN in Time Series Space				
Toe1	**Toe2**	**FordA**	**FordB**	**compared with**
0.3200	0.1920	0.3410	0.4420	ED
0.2280	0.1620	0.4380	0.4060	DTW

SVM in Dissimilarity Space				
Toe1	**Toe2**	**FordA**	**FordB**	**embedded with**
0.1404	0.1615	0.2885	0.2990	A = Training Set
0.1140	0.2231	0.2458	0.2624	B = Complementary Set
0.0877	0.1923	0.2438	0.2525	AB
0.1930	0.1769	0.3396	0.3707	C = Ham
0.2632	0.3385	0.4135	0.4497	D = Herring
0.1798	0.1692	0.3291	0.3804	CD
0.0877	0.2000	0.2413	0.2486	ABCD
0.2675	0.3231	0.4452	0.4568	E = Coffee
0.1316	0.2000	0.3107	0.3328	F = Wafer
0.1711	0.2077	0.3035	0.3342	EF
0.1535	0.1769	0.2399	0.2371	ABEF
0.1447	0.1846	0.2291	0.2368	ABCDEF
0.2237	0.2923	0.3213	0.3377	G = Random Process 1
0.2719	0.1769	0.4263	0.4054	H = Random Process 2
0.1930	0.2385	0.3121	0.3276	GH
0.1491	0.2231	0.2363	0.2428	ABGH
0.1535	0.1923	0.2205	0.2280	ABCDEFGH

[1] Set G and H each contain 500 random sequences that were generated by an auto-regressive progress $x_i = ax_{i-1} - bx_{i-2} + c\eta_i$ of second order, which was initialized with $x_1 = x_2 = 0$ and stop after 500 time steps. The parameter settings for G and H are $a = 1.8, b = 0.972, c = 0.64$ and $a = 1.85, b = 0.917, c = 0.76$ respectively. The noise η is a vector of uniformly distributed random numbers in the interval $(0, 1)$

4 Discussion

Having explained our approach and experimental setup, we are eventually in the position to discuss the empirical results presented in Table 2. In the following Table 3 we compare the classification errors of (i) the traditional 1-Nearest-Neighbor approach using either ED or DTW as competitor - *1NN Baseline* [4], (ii) the standard dissimilarity spaces approach using only training examples for embedding - *DSS Standard* [7, 8], and (iii) our proposed transfer learning in dissimilarity spaces approach using the best combination of constructed feature vectors - *DSS Transfer*.

Table 3. Comparison of Classification Errors and Performance Increase

Toe1	Toe2	FordA	FordB	
0.2280	0.1620	0.3410	0.4060	(i) 1NN Baseline
0.1404 (+ 38.42%)	0.1615 (+ 0.31%)	0.2885 (+ 15.41%)	0.2990 (+ 26.36%)	(ii) DSS Standard
0.0877 (+ 61.54%)	0.1692 (- 4.44%)	0.2205 (+ 35.34%)	0.2280 (+ 43.84%)	(iii) DSS Transfer

As shown in Table 3, for the *Toe1* dataset the *DSS Standard* approach achieved a performance increase of about 38%, while our proposed *DSS Transfer* approach even achieved a performance increase of more than 61% (relative to the *1NN Baseline*). In the case of *Toe1*, the lowest classification error was given by combining the feature vectors of dissimilarity spaces A and B, which were constructed from the corresponding $Toe1$ training examples and the complementary *Toe2* training and test set (see Table 2).

Our proposed *DSS Transfer* approach furthermore outperformed the other techniques for the *FordA* and *FordB* dataset. For these two datasets, the lowest classification error was given by combining the feature vectors of all examined dissimilarity spaces $A - H$ (refer to Table 2). However, the *DSS* (*Standard* and *Transfer*) approach were not able to achieve a performance increase for the *Toe2* dataset, which may be due to the already quite small *1NN Baseline* classification error (for *Toe2*).

In general, the results in Table 2 show that transfer learning from arbitrary or even random dissimilarity spaces is often able to achieve significantly higher classification accuracy than learning from the original training examples. This is astonishing since arbitrary or random dissimilarity representations contain no domain knowledge that relates to the original classification problem. In that sense, transfer learning can be imagined as solving the classification of cats and dogs by means of knowledge about pigs and cows or random mammals.

For instance, Table 2 shows that in the case of *Toe1* our *DSS Transfer* approach achieved a classification error of 0.1140 using only knowledge about the complementary *Toe2* set, which equates to a performance increase of exactly 50% with respect to the *1NN Baseline* approach using all of the available domain knowledge. Furthermore, in the case of *FordB* our *DSS Transfer* approach achieved a classification error of 0.3276 using a combination of random sequences

from set G and H, which is a performance increase of more than 19% in comparison to the *1NN Baseline* approach. Of course, there are also time series datasets, such as *Herring*, which yield dissimilarity representations that result in a performance decrease (for all considered classification problems).

In the following, we discuss how the dimension of the dissimilarity spaces influences the classification accuracy. Table 1 shows that the dimension of our individual dissimilarity spaces correlates with the number of time series that were used for embedding. The question is whether more data leads to better dissimilarity representations and lower classification error? Figure 1 illustrates the classification accuracy for *Toe1* and *Toe2* as a function of the dataset size.

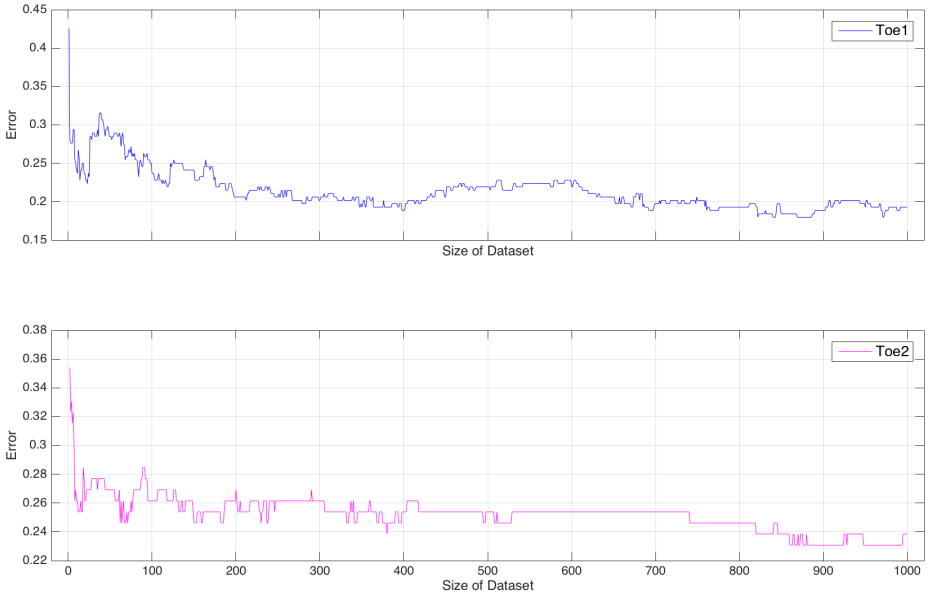

Fig. 1. Classification error for *Toe1* and *Toe2* with varying dataset size, where the time series for the dissimilarity embedding were randomly selected from set G and H.

According to our experimental results shown in Figure 1, the classification error converges to a certain lower limit with growing dataset size or embedding dimension. However, the error curves are strongly fluctuating, meaning that in certain cases additional information worsens the classification accuracy and, therefore, more data is not necessarily better. Earlier work [7, 8] suggested to employ dimensionality reduction techniques to identify those time series that add useful information to the dissimilarity representation.

5 Conclusion

We have picked up the idea of time series classification in dissimilarity spaces [7, 8] and extended this approach to transfer learning, which allows us to generate enriched dissimilarity representations by considering additional time series that are totally unrelated to the original supervised learning task. Our results show that the proposed approach is able to considerately improve classification accuracy, depending on the additional time series used for the embedding.

In general, it would be advantageous if we could determine the information gain of additional time series datasets beforehand. In future work, we aim to investigate the influence of the (original and transfer) data distribution on the performance increase achieved by the enriched dissimilarity representations.

References

1. Batista GE, Keogh EJ, Tataw OM, De Souza VMA: *CID: an efficient complexity-invariant distance for time series.* Data Mining and Knowledge Discovery, 2013.
2. Bunke H, Riesen K: *Graph classification based on dissimilarity space embedding.* Structural, Syntactic, and Statistical Pattern Recognition, LNCS, 2008.
3. Cao LJ, Chua KS, Chong WK, Lee HP, Gu QM: *A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine.* Neurocomp., 2003.
4. Chen Y, Keogh EJ, Hu B, Begum N, Bagnall A, Mueen A, Batista GE: *The UCR Time Series Archive.* `www.cs.ucr.edu/~eamonn/time_series_data/`
5. Duin RPW, Pekalska E: *The dissimilarity space: Bridging structural and statistical pattern recognition.* Pattern Recognition Letter, 2011.
6. Hofstadter D, Sander E: *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking.* Basic Books, 2013.
7. Jain BJ, Spiegel S: *Time Series Classification in Dissimilarity Spaces.* Advanced Analytics and Learning on Temporal Data, ECML-PKDD, 2015.
8. Jain BJ, Spiegel S: *Dimension Reduction in Dissimilarity Spaces for Time Series Classification.* Advanced Analysis and Learning on Temporal Data, LNAI, 2016.
9. Kate RJ: *Using dynamic time warping distances as features for improved time series classification.* Data Mining and Knowledge Discovery, 2015.
10. Lines J, Bagnall A: *Time series classification with ensembles of elastic distance measures.* Data Mining and Knowledge Discovery, 2014.
11. Pan SJ, Yang Q: *A Survey on Transfer Learning.* IEEE Transactions on Knowledge and Data Engineering, 2010.
12. Pekalska E, Duin RPW: *The Dissimilarity Representation for Pattern Recognition.* World Scientific Publishing Co., Inc., 2005.
13. Raina R, Battle A, Lee H, Packer B, Ng AY: *Self-taught learning: transfer learning from unlabeled data.* International Conference on Machine Learning, 2007.
14. Rakthanmanon T, Campana B, Mueen A, Batista GE, Westover B, Zhu Q, Zakaria J, Keogh EJ: *Mining trillions of time series subsequences under dynamic time warping.* ACM Transaction on Knowledge Discovery from Data, 2013.
15. Riesen K, Neuhaus M, Bunke H: *Graph embedding in vector spaces by means of prototype selection.* Graph-based Representations in Pattern Recognition, 2007.
16. Subasi A, Gursoy MI: *EEG signal classification using PCA, ICA, LDA and support vector machines.* Expert System with Applications, 2010.
17. Xi X, Keogh EJ, Shelton C, Wei L, Ratanamahatana CA: *Fast time series classification using numerosity reduction.* Int. Conf. on Machine Learning, 2006.